numero 222
febbraio 1987
anno XX
volume XXXVIII

LE SCIENZE
edizione italiana di SCIENTIFIC AMERICAN

Il caos

Vi è ordine nel caos: l'aleatorietà possiede una forma geometrica soggiacente; il caos pone limitazioni fondamentali alla prevedibilità, ma suggerisce relazioni causali dove prima non parevano possibili

di James P. Crutchfield, J. Doyne Farmer, Norman H. Packard e Robert S. Shaw

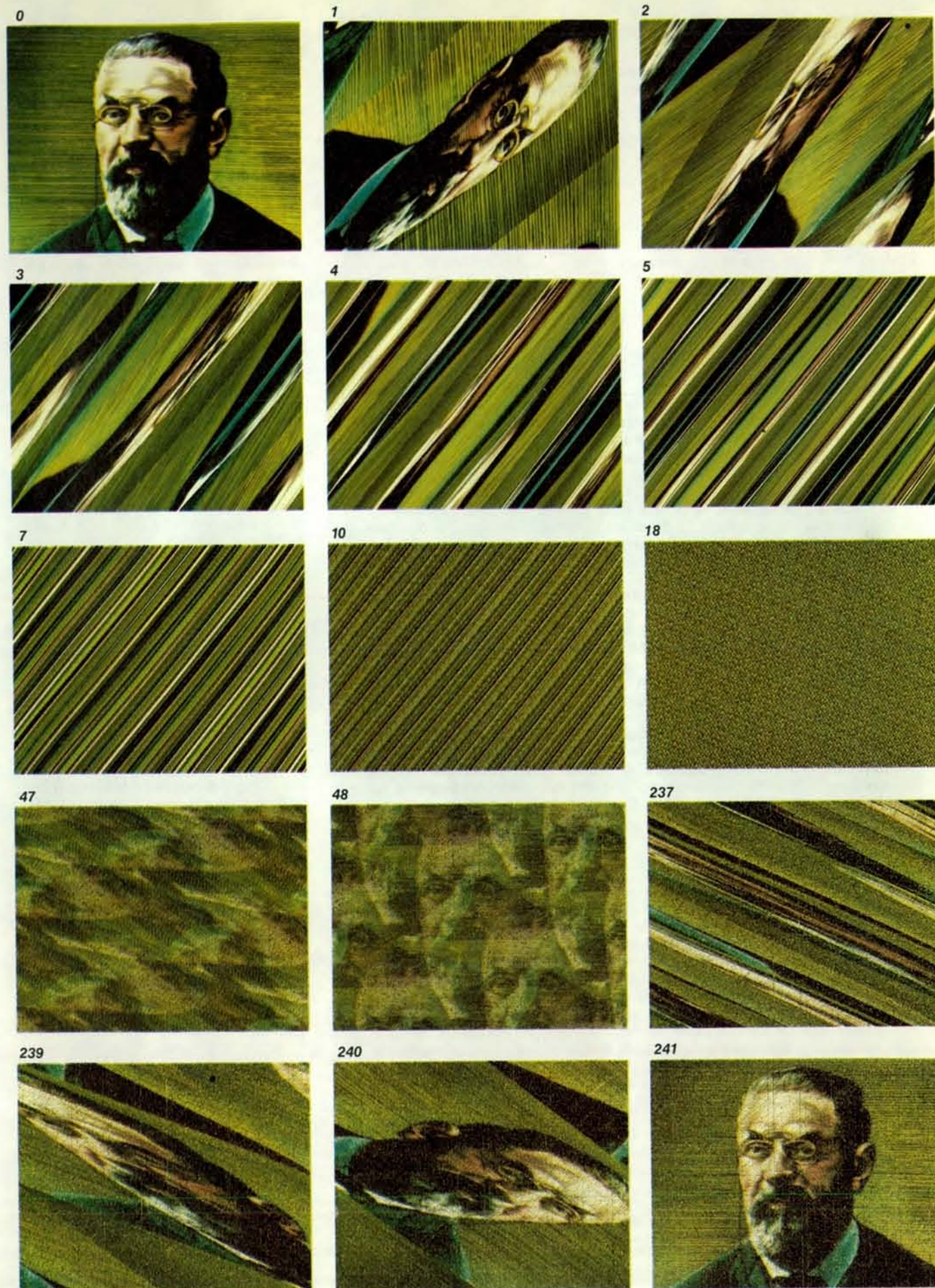
La grande forza della scienza sta nella sua capacità di collegare cause ed effetti. Sulla base delle leggi della gravitazione, per esempio, è possibile prevedere le eclissi con migliaia di anni di anticipo. Altri fenomeni naturali, però, non sono altrettanto prevedibili. Il moto dell'atmosfera segue le leggi della fisica al pari del moto dei pianeti; tuttavia le previsioni del tempo vengono ancora espresse in termini probabilistici. L'evoluzione delle condizioni meteorologiche, lo scorrere di un ruscello, il rotolare di un dado presentano aspetti imprevedibili. Poiché in questi fenomeni non esiste una chiara relazione tra causa ed effetto, si dice che contengono elementi aleatori. Eppure fino a poco tempo fa non vi erano validi motivi per dubitare che in linea di principio si potesse giungere a una prevedibilità rigorosa. Si riteneva che bastasse raccogliere ed elaborare una quantità sufficiente di informazione.

Questa opinione è stata modificata da una scoperta sorprendente. Semplici sistemi deterministici, anche costituiti da pochi elementi, possono manifestare un comportamento aleatorio. Questa aleatorietà è di natura essenziale e non scompare se si raccolgono ulteriori informazioni. A questo genere di aleatorietà si è dato il nome di caos.

Un paradosso apparente è che il caos è deterministico, cioè è generato da regole fisse che, di per sé, non contengono alcun elemento casuale. In linea di principio il futuro è determinato completamente dal passato, ma in pratica le piccole indeterminazioni vengono amplificate; quindi, benché il comportamento sia prevedibile a breve scadenza, alla lunga risulta imprevedibile. Nel caos vi è ordine: soggiacenti al comportamento caotico vi sono eleganti forme geometriche che creano l'aleatorietà così come il cartaio mescola un mazzo di carte o un cuoco mescola l'impasto di un dolce.

La scoperta del caos ha creato un paradigma nuovo fra i modelli scientifici. Da una parte comporta l'esistenza di nuove limitazioni fondamentali alla nostra capacità di compiere previsioni; dall'altra, il determinismo inerente al caos implica che molti fenomeni aleatori sono più prevedibili di quanto si pensasse. Informazioni apparentemente aleatorie raccolte in passato (e archiviate perché ritenute troppo complicate) oggi possono essere spiegate in termini di leggi semplici. Il caos ci consente di scoprire l'ordine in sistemi diversissimi fra loro come l'atmosfera, un rubinetto che gocciola e il cuore. Ne è conseguita una rivoluzione che sta coinvolgendo molte branche diverse della scienza.

Che origine ha il comportamento caotico? Il moto browniano è un esempio classico di aleatorietà. Osservando al microscopio un granello di polvere in una bacinella d'acqua, lo si vede muoversi con un tremolio continuo e capriccioso. Ciò è dovuto alle molecole d'acqua circostanti che, mosse dall'agitazione termica, bombardano la particella di polvere. Poiché le molecole d'acqua sono invisibili e sono numerosissime, il movimento preciso della particella di polvere è affatto imprevedibile. Qui la rete delle azioni causali tra le varie sottounità può essere così intricata da rendere affatto aleatoria la configurazione del moto risultante.

Il caos deriva dall'operazione geometrica dello stiramento. Su un ritratto del matematico francese Henri Poincaré, il fondatore della teoria dei sistemi dinamici, ne è illustrato l'effetto. L'immagine di partenza (*in alto a sinistra*) è stata discretizzata per consentire a un calcolatore di effettuare lo stiramento. Una semplice trasformazione matematica stira l'immagine diagonalmente, come se fosse dipinta su un foglio di gomma. Quando il foglio esce dal riquadro, viene tagliato e reinserito dall'altra parte, come si vede nell'immagine *1*. (Il numero sopra ciascuna immagine indica quante volte è stata effettuata la trasformazione.) La ripetizione dell'operazione ha per effetto di scompaginare la faccia (*immagini* 2-4). Il risultato finale è una combinazione casuale di colori, che si presenta come uno strato omogeneo di verde (*immagini* 10 *e* 18). Talvolta accade che alcuni punti tornino in prossimità delle loro posizioni iniziali, provocando una breve ricomparsa dell'immagine originale (*immagini* 47-48, 239-241). La trasformazione illustrata è particolare in quanto il fenomeno della «ricorrenza di Poincaré» (come viene chiamata in meccanica statistica) si presenta molto più spesso del solito; in generale in una trasformazione caotica generica la ricorrenza è rarissima e si presenta forse una volta sola nel corso della vita dell'universo. In presenza di fluttazioni di fondo di entità qualunque, l'intervallo tra una ricorrenza e l'altra è di solito così lungo che tutta l'informazione relativa all'immagine originale va perduta.

Il tipo di caos di cui parleremo non ha bisogno di un gran numero di sottounità o di azioni invisibili. La presenza di un comportamento aleatorio in sistemi semplicissimi giustifica un riesame delle fonti dell'aleatorietà anche in grandi sistemi, come quello meteorologico.

Che cosa rende il moto dell'atmosfera tanto più difficile da prevedere del moto del sistema solare? Entrambi i sistemi sono costituiti da molte componenti e sono retti dalla seconda legge della dinamica, $F = ma$, che può essere considerata una semplice ricetta per prevedere il futuro. Se le forze F agenti su una massa m assegnata sono note, lo è anche l'accelerazione a. Dalle regole dell'analisi segue allora che se la posizione e la velocità di un oggetto possono essere misurate a un istante dato, esse restano determinate per sempre. Si tratta di un principio molto potente, tanto che nel Settecento il grande matematico francese Simon de Laplace affermò che, date la posizione e la velocità di tutte le particelle dell'universo, egli sarebbe stato capace di prevederne l'evoluzione futura per l'eternità. Benché molte evidenti difficoltà pratiche si oppongano alla realizzazione del programma di Laplace, per oltre un secolo sembrò che nulla ne infirmasse la legittimità, almeno in linea di principio. L'applicazione letterale dell'affermazione di Laplace al comportamento umano portò alla conclusione filosofica che questo comportamento fosse completamente predeterminato e che il libero arbitrio non esistesse.

Laplace, 1776

«Lo stato attuale del sistema della natura consegue evidentemente da quello che era all'istante precedente e se noi immaginassimo un'intelligenza che a un istante dato comprendesse tutte le relazioni fra le entità di questo universo, essa potrebbe conoscere le rispettive posizioni, i moti e le disposizioni generali di tutte quelle entità in qualunque istante del passato o del futuro.

«L'astronomia fisica, la branca della scienza che rende il massimo onore alla mente umana, ci dà un'idea, ancorché imperfetta, di quale sarebbe questa intelligenza. La semplicità della legge che regola il moto dei corpi celesti e le relazioni fra le loro masse e distanze ci consentono di seguire i loro moti con l'analisi fino a un certo punto; e per poter determinare lo stato del sistema di questi grandi corpi nei secoli passati o in quelli futuri è sufficiente al matematico conoscere tramite l'osservazione la loro posizione e la loro velocità in un istante qualunque. L'uomo deve questo vantaggio alla potenza degli strumenti impiegati e al piccolo numero di relazioni che entrano nei suo calcoli. Ma l'ignoranza delle diverse cause che concorrono alla formazione degli eventi come pure la loro complessità, insieme con l'imperfezione dell'analisi, ci impediscono di conseguire la stessa certezza rispetto alla grande maggioranza dei fenomeni. Vi sono quindi cose che per noi sono incerte, cose più o meno probabili, e noi cerchiamo di rimediare all'impossibilità di conoscerle determinando i loro diversi gradi di verosimiglianza. Accade così che alla debolezza della mente umana si debba una delle più fini e ingegnose fra le teorie matematiche, la scienza del caso o della probabilità.»

Poincaré, 1903

«Una causa piccolissima che sfugga alla nostra attenzione determina un effetto considerevole che non possiamo mancar di vedere, e allora diciamo che l'effetto è dovuto al caso. Se conoscessimo esattamente le leggi della natura e la situazione dell'universo all'istante iniziale, potremmo prevedere esattamente la situazione dello stesso universo in un istante successivo. Ma se pure accadesse che le leggi naturali non avessero più alcun segreto per noi, anche in tal caso potremmo conoscere la situazione iniziale solo *approssimativamente*. Se questo ci permettesse di prevedere la situazione successiva con la stessa *approssimazione*, non ci occorrerebbe di più e dovremmo dire che il fenomeno è stato previsto, che è governato da leggi. Ma non sempre è così; può accadere che piccole differenze nelle condizioni iniziali ne producano di grandissime nei fenomeni finali. Un piccolo errore nelle prime produce un errore enorme nei secondi. La previsione diviene impossibile e si ha un fenomeno fortuito.»

Le opinioni di due luminari sul caso e sulla probabilità messe a confronto. Il matematico francese Pierre Simon de Laplace sosteneva che le leggi della natura implicano un rigido determinismo e una prevedibilità totale, benché le imperfezioni delle osservazioni rendano necessario introdurre la teoria delle probabilità. Nella citazione di Poincaré è prefigurata l'opinione odierna secondo al quale indeterminazioni arbitrariamente piccole sullo stato del sistema possono amplificarsi con il passare del tempo, rendendo impossibili le previsioni sul futuro lontano.

La scienza del Novecento ha visto il crollo del determinismo di Laplace e per due ragioni molto diverse. La prima è la meccanica quantistica: un dogma fondamentale di questa teoria è il principio d'indeterminazione di Heisenberg, secondo il quale vi è una limitazione essenziale alla precisione con cui possono essere misurate la posizione e la velocità di una particella. Questa indeterminazione dà una buona spiegazione di certi fenomeni aleatori, come la disintegrazione radioattiva. Il nucleo è così piccolo che il principio d'indeterminazione pone una limitazione fondamentale alla conoscenza del suo moto ed è impossibile raccogliere informazioni sufficienti a prevedere il momento della sua disintegrazione.

Su scala più grande, però, la fonte dell'imprevedibilità dev'essere cercata altrove. Alcuni fenomeni macroscopici sono prevedibili e altri no, e questa distinzione non ha nulla a che fare con la meccanica quantistica. La traiettoria di una palla, per esempio, è intrinsecamente prevedibile, e un giocatore sfrutta istintivamente questo fatto. Invece la traiettoria di un palloncino spinto dall'aria che ne esce impetuosamente non è prevedibile; il palloncino caracolla e muta rotta capricciosamente in istanti e in punti che è assolutamente impossibile prevedere. Il palloncino segue le leggi di Newton esattamente come la palla: allora, perché è tanto più difficile prevederne il comportamento?

L'esempio classico di questa dicotomia è il moto dei fluidi. In certe condizioni il moto di un fluido è laminare (cioè uniforme, stazionario e regolare) e facile da prevedere a partire dalle equazioni. In altre condizioni il moto del fluido è turbolento (cioè non uniforme, non stazionario e irregolare) e quindi difficile da prevedere. Il passaggio dal comportamento laminare a quello turbolento è ben noto a chiunque, viaggiando in aereo, sia passato improvvisamente da una zona di calma a una di tempesta. Qual è la causa della differenza essenziale tra il moto laminare e quello turbolento?

Per capire appieno perché si tratti di un vero e proprio enigma, pensiamo a un ruscello di montagna. L'acqua turbina e schizza come se avesse una mente propria, spostandosi prima di qua e poi di là. Tuttavia i massi sono ben piantati nel letto del ruscello e i suoi tributari gli forniscono acqua a un tasso quasi costante. Da dove nasce allora l'aleatorietà del moto dell'acqua?

Al fisico sovietico Lev D. Landau, scomparso di recente, è attribuita una spiegazione del moto aleatorio dei fluidi che ha tenuto il campo per molti anni: il moto di un fluido turbolento conterrebbe molte oscillazioni diverse e indipendenti. Via via che il fluido si muove più in fretta, e diventa perciò più turbolento, le oscillazioni entrano nel moto una alla volta. Benché le singole oscillazioni possano essere semplici, la complessità del moto combinato rende impossibile prevedere il flusso complessivo.

La teoria di Landau, tuttavia, è stata confutata. Il comportamento aleatorio si presenta anche in sistemi semplicissimi, senza bisogno di complicazioni o indeterminazioni. Il matematico francese Henri Poincaré se ne rese conto all'inizio del secolo, quando osservò che fenomeni imprevedibili e «fortuiti» possono presentarsi nei sistemi in cui una piccola variazione nel presente provoca una variazione molto più ampia nel futuro. Il concetto si chiarisce se si pensa a un sasso posto in equilibrio sulla cima di un monte. Una spintarella in un verso o in un altro basta a farlo rotolare giù lungo percorsi molto diversi. Il sasso è sensibile a piccole azioni solo quando è sulla cima del monte; i sistemi caotici invece sono sensibili in ogni punto del loro moto.

Lo spazio degli stati è un concetto utile per visualizzare il comportamento di un sistema dinamico. Si tratta di uno spazio astratto le cui coordinate sono i gradi di libertà del moto del sistema. Il moto di un pendolo (*in alto*), per esempio, è determinato dalla posizione e dalla velocità iniziali. Il suo stato è quindi un punto di un piano le cui coordinate sono posizione e velocità (*in basso*). Quando il pendolo compie le sue oscillazioni il punto descrive un'«orbita», cioè un cammino, nello spazio degli stati. Per un pendolo ideale, senza attrito, l'orbita è una curva chiusa (*in basso a sinistra*); se c'è attrito, l'orbita è una spirale che tende a un punto (*in basso a destra*).

Un semplice esempio servirà a illustrare quanto certi sistemi fisici siano sensibili alle azioni esterne. Consideriamo una partita a biliardo; idealizzandola un po' si può supporre che le palle si muovano sul tavolo e si urtino con una perdita di energia trascurabile. Con un solo colpo il giocatore provoca fra tutte le palle una prolungata successione di collisioni. Naturalmente il giocatore desidera conoscere gli effetti del colpo. Se possedesse un controllo perfetto del proprio colpo, per quanto tempo potrebbe dimostrarsi valida la previsione della traiettoria della palla colpita dalla stecca? Ebbene, se il giocatore ignorasse effetti anche minimi, per esempio l'attrazione gravitazionale di un elettrone situato ai margini della galassia, la sua previsione sarebbe errata già dopo un solo minuto!

Questo cospicuo aumento dell'indeterminazione deriva dal fatto che le palle sono curve e che a ciascuna collisione vengono amplificate le piccole differenze nel punto di contatto. Quest'amplificazione è esponenziale: viene composta a ogni collisione, come accade per le generazioni successive dei batteri quando spazio e cibo siano illimitati. Qualunque effetto, per quanto piccolo, acquista rapidamente proporzioni macroscopiche. Questa è una delle caratteristiche fondamentali del caos.

L'amplificazione esponenziale degli errori dovuta alla dinamica del caos costituisce la seconda ragione dell'insuccesso di Laplace. La meccanica quantistica comporta che le misure iniziali siano sempre imprecise e il caos fa sì che le imprecisioni superino rapidamente la capacità di fare previsioni. Senza il caos Laplace avrebbe potuto sperare che gli errori restassero limitati, o almeno crescessero abbastanza lentamente da consentirgli di fare previsioni a lungo termine. In presenza del caos le previsioni sono invece destinate a raggiungere ben presto un'imprecisione madornale.

Il più ampio quadro concettuale dal quale il caos emerge è la cosiddetta teoria dei sistemi dinamici. Un sistema dinamico si compone di due parti: le caratteristiche del suo stato (cioè le informazioni essenziali sul sistema) e la dinamica (una regola che descrive l'evoluzione dello stato nel tempo). L'evoluzione può essere visualizzata in uno spazio degli stati, una costruzione astratta le cui coordinate sono le componenti dello stato. In generale le coordinate dello spazio degli stati variano secondo il contesto; per un sistema meccanico potrebbero essere la posizione e la velocità, ma per un modello ecologico potrebbero essere le popolazioni delle diverse specie.

Un buon esempio di sistema dinamico è offerto dal pendolo semplice. Per descrivere completamente il suo moto bastano due variabili: la posizione e la velocità. Lo stato è quindi un punto del piano le cui coordinate sono posizione e velocità. Le leggi di Newton forniscono una regola, che assume la forma matematica di un'equazione differenziale, per descrivere l'evoluzione dello stato. Quando il pendolo oscilla avanti e indietro, il suo stato si sposta lungo una curva del piano, detta «orbita». Nel caso ideale di un pendolo privo di attrito, l'orbita è una curva chiusa; altrimenti è una curva a spirale che converge verso un punto corrispondente allo stato di quiete del pendolo.

L'evoluzione temporale di un sistema dinamico può avvenire sia in tempo continuo sia in tempo discreto. Nel primo caso si parla di flusso, nel secondo di applicazione. Un pendolo si muove con continuità da uno stato a un altro e quindi viene descritto da un flusso in tempo continuo. Il numero degli insetti che nascono ogni anno in una data zona e l'intervallo temporale tra le gocce di un rubinetto che gocciola vengono invece descritti in modo più naturale da un'applicazione in tempo discreto.

Per vedere come un sistema si evolva a partire da un dato stato iniziale si può usare la dinamica (cioè le equazioni del moto) per effettuare uno spostamento incrementale lungo un'orbita. Questo metodo per ricavare il comportamento del sistema richiede un impegno di calcolo proporzionale all'intervallo di tempo in cui si desidera seguire l'orbita. Per certi sistemi semplici, come il pendolo senza attrito, le equazioni del moto possono talora possedere una soluzione in forma chiusa, che è una formula che esprime qualunque stato futuro in termini dello stato iniziale. Una soluzione in forma chiusa rappresenta una scorciatoia, un algoritmo più semplice, che consente di prevedere il futuro solo in base allo stato iniziale e all'istante finale, senza bisogno di passare per gli stati intermedi. Con una soluzione di questo tipo l'impegno algoritmico necessario per seguire il moto del sistema è grosso modo indipendente dal tempo desiderato. Date le equazioni del moto dei pianeti e della Luna, nonché le posizioni e le velocità della Terra e della Luna, si possono prevedere, per esempio, le eclissi con anni di anticipo.

I successi ottenuti agli inizi della fisica nel trovare soluzioni in forma chiusa per svariati sistemi semplici fecero nascere la speranza che soluzioni di questo genere esistessero per qualunque sistema meccanico. Oggi si sa che ciò, purtroppo, in

generale non è vero. Il comportamento imprevedibile dei sistemi dinamici caotici non può essere espresso mediante una soluzione in forma chiusa e di conseguenza non esiste alcuna scorciatoia per prevedere tale comportamento.

Nondimeno lo spazio degli stati costituisce uno strumento potente per descrivere il comportamento dei sistemi caotici. Lo spazio degli stati è utile perché consente di rappresentare il comportamento in forma geometrica. Per esempio, un pendolo soggetto ad attrito prima o poi si arresta e ciò nello spazio degli stati corrisponde a un'orbita che tende verso un punto. Il punto non si muove - è un punto fisso - e, poiché attrae le orbite vicine, lo si chiama attrattore. Se si dà una debole spinta a un pendolo, esso torna allo stesso attrattore fisso. Tutti i sistemi che col tempo pervengono a uno stato di quiete possono essere caratterizzati da un punto fisso nello spazio degli stati. Questo è un caso particolare di un fenomeno generalissimo, che si presenta quando le perdite, dovute per esempio all'attrito o alla viscosità, fanno sì che le orbite siano attratte verso una regione più piccola dello spazio degli stati avente un numero inferiore di dimensioni. Anche queste regioni sono chiamate attrattori. In termini grossolani, un attrattore è ciò in cui si stabilizza o verso cui è attratto il comportamento di un sistema.

Alcuni sistemi alla lunga non tendono alla quiete e passano invece ciclicamente per una successione periodica di stati. Un esempio è costituito dall'orologio a pendolo, in cui l'energia perduta per gli attriti viene reintegrata da una molla o da un peso. Il pendolo continua a ripetere lo stesso moto. Nello spazio degli stati un moto siffatto corrisponde a un'orbita periodica o ciclo. Indipendentemente da come un pendolo venga posto in oscillazione, il ciclo che rappresenta il limite cui esso tende a lungo termine è sempre lo stesso. Gli attrattori di questo tipo si chiamano pertanto cicli limite. Un altro sistema ben noto che ha come attrattore un ciclo limite è il cuore.

Un sistema può possedere più attrattori. In tal caso condizioni iniziali diverse possono portare ad attrattori diversi. L'insieme dei punti che si evolvono verso un dato attrattore costituisce il suo bacino di attrazione. L'orologio a pendolo possiede due di questi bacini: se lo spostamento del pendolo dalla posizione di quiete è piccolo, esso torna in questa posizione; se invece lo spostamento è grande, l'orologio si avvia e il pendolo si mette a oscillare in modo stabile.

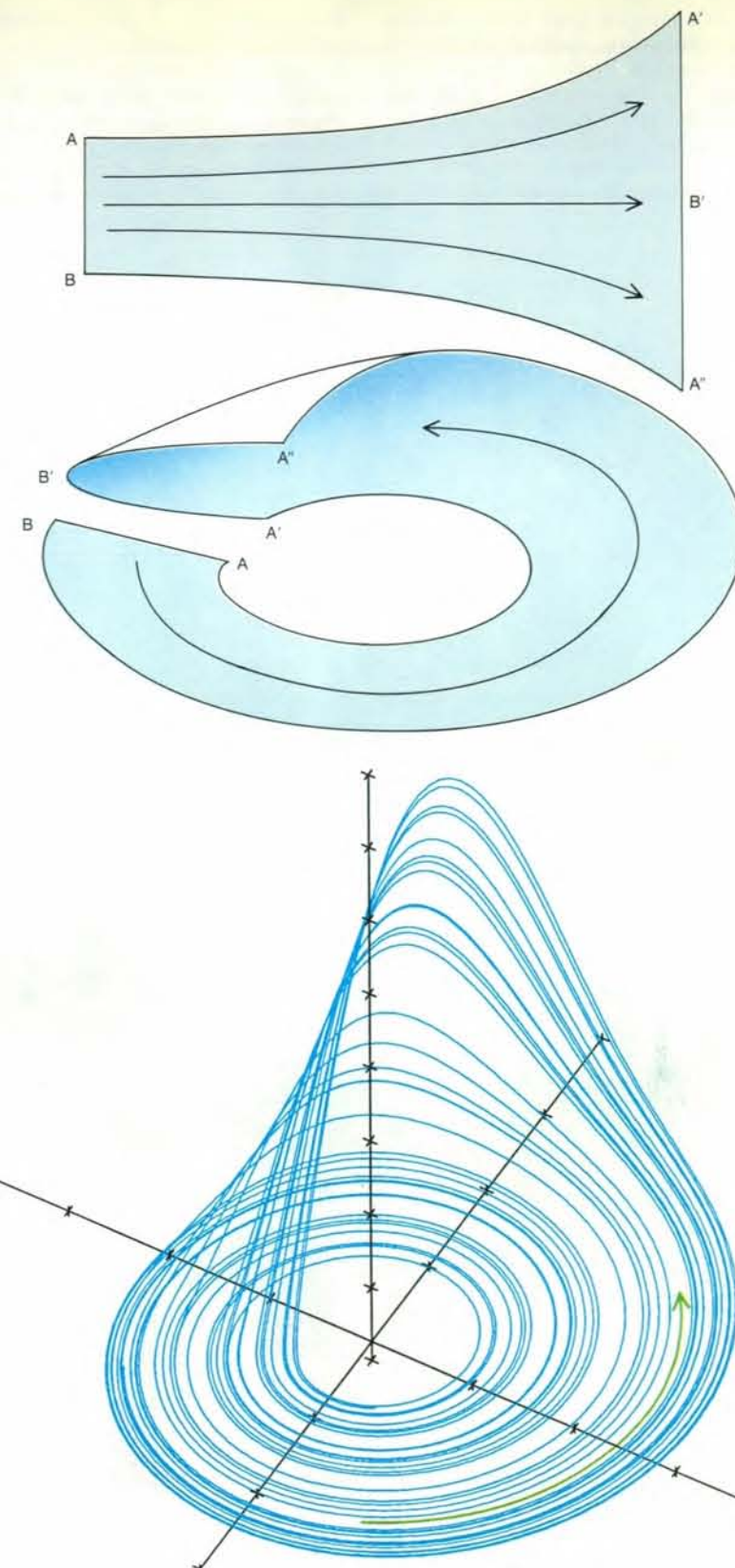
La forma di attrattore immediatamente più complicata è il toro, che somiglia alla superficie di una ciambella. Questa forma descrive un moto costituito da due oscillazioni indipendenti, moto talvolta detto quasiperiodico. (Se ne possono costruire esempi fisici mediante oscillatori elettrici pilotati.) Nello spazio degli stati l'orbita si avvolge intorno al toro; una delle due frequenze è determinata dalla velocità con cui l'orbita gira intorno al toro nella dimensione minore, mentre l'altra dipende dalla velocità con cui l'orbita si sposta lungo la dimensione maggiore. Gli attrattori possono essere anche tori con un numero più elevato di dimensioni e che rappresentano una combinazione di più di due oscillazioni.

La caratteristica più importante del moto quasiperiodico è che, nonostante la sua complessità, esso è prevedibile. Anche se l'orbita non si ripete mai esattamente, come accade se le frequenze del moto non hanno un divisore comune, il moto resta regolare. Le orbite che cominciano vicine tra loro sul toro rimangono vicine tra loro e la prevedibilità a lungo termine è assicurata.

Gli attrattori sono forme geometriche che caratterizzano il comportamento a lungo termine nello spazio degli stati. In altri termini, un attrattore è ciò verso cui si stabilizza, o è attratto, il comportamento di un sistema. Nel caso raffigurato gli attrattori sono in blu mentre gli stati iniziali sono in rosso. Le traiettorie (*in verde*), partendo dagli stati iniziali, finiscono con l'avvicinarsi agli attrattori. Il tipo di attrattore più semplice è un punto fisso (*in alto a sinistra*). Questo attrattore corrisponde a un pendolo soggetto ad attrito: il pendolo ritorna sempre nella stessa posizione di quiete, indipendentemente da come abbia cominciato a oscillare (*si veda la parte a destra dell'illustrazione della pagina precedente*). L'attrattore immediatamente più complicato è un ciclo limite (*in alto al centro*), che nello spazio degli stati è una curva chiusa. Un ciclo limite descrive oscillazioni stabili, come il moto di un orologio a pendolo o il battito cardiaco. Le oscillazioni composte, caratteristiche del comportamento quasiperiodico, corrispondono a un attrattore a forma di toro (*in alto a destra*). Tutti e tre questi attrattori sono prevedibili: il loro comportamento può essere previsto con una precisione arbitraria. Gli attrattori caotici, invece, corrispondono a moti imprevedibili e hanno una forma geometrica più complicata. Nella parte inferiore dell'immagine sono illustrati tre esempi di attrattori caotici; da sinistra a destra, essi sono stati scoperti rispettivamente da Edward N. Lorenz, da Otto E. Rössler e da uno degli autori (Shaw). Le immagini sono state ottenute ricorrendo a semplici sistemi di equazioni differenziali aventi uno spazio degli stati tridimensionale.

Fino a tempi piuttosto recenti gli unici attrattori conosciuti erano i punti, i cicli limite e i tori. Nel 1963 Edward N. Lorenz, del Massachusetts Institute of Technology, scoprì un esempio concreto di sistema a poche dimensioni che ha un comportamento complesso. Spinto dal desiderio di comprendere l'imprevedibilità delle condizioni meteorologiche, Lorenz partì dalle equazioni del moto di un fluido (l'atmosfera può essere considerata un fluido) e semplificandole ottenne un sistema dotato di tre soli gradi di libertà. Il sistema, però, si comportava in un modo evidentemente aleatorio che non poteva essere caratterizzato adeguatamente da nessuno dei tre attrattori noti a quel tempo. L'attrattore che osservò, e che ora viene chiamato attrattore di Lorenz, fu il primo esempio di attrattore caotico, o attrattore strano.

Servendosi di un calcolatore digitale per simulare il proprio semplice modello, Lorenz chiarì il meccanismo fondamentale che dava luogo all'aleatorietà osservata: le perturbazioni microscopiche vengono amplificate fino a interferire con il comportamento macroscopico. Due orbite corrispondenti a condizioni iniziali prossime divergono con velocità esponenziale e quindi restano vicine tra loro soltanto per breve tempo. Per gli attrattori non caotici la situazione è qualitativamente diversa, poiché le orbite vicine restano vicine l'una all'altra, gli errori piccoli restano limitati e il comportamento è prevedibile.

La chiave per interpretare il comportamento caotico sta nella comprensione di una semplice operazione di stiramento e piegatura, che ha luogo nello spazio degli stati. La divergenza esponenziale è un fenomeno locale: dal momento che la dimensione (*size*) degli attrattori è finita, due orbite situate su un attrattore caotico non possono continuare a divergere esponenzialmente per sempre. Ne segue che l'attrattore deve ripiegarsi su se stesso. Benché le orbite divergano e seguano strade sempre più diverse, prima o poi devono passare di nuovo una accanto all'altra. Le orbite situate su un attrattore caotico vengono mescolate da questo processo più o meno come un mazzo di carte viene mescolato dal car-

Un attrattore caotico ha una struttura molto più complicata di uno prevedibile come un punto, un ciclo limite o un toro. Se molto ingrandito, un attrattore caotico si presenta come una superficie non regolare e contenente pieghe. L'immagine presenta i passaggi che portano alla costruzione di un attrattore caotico nel caso più semplice: quello di Rössler (*in basso*). Le traiettorie vicine sull'oggetto devono «stirarsi», cioè divergere, esponenzialmente (*in alto*); nel caso illustrato la distanza fra traiettorie vicine diventa circa il doppio. Inoltre, per restare compatto, l'oggetto deve «ripiegarsi» su se stesso (*al centro*): la superficie si piega su se stessa finché le due estremità s'incontrano. Osservato in molti sistemi, dal moto dei fluidi alle reazioni chimiche, l'attrattore di Rössler illustra la massima di Einstein che la natura predilige le forme semplici.

taio. L'aleatorietà delle orbite caotiche è una conseguenza di questo processo di mescolamento. Il processo di stiramento e piegatura avviene più volte e produce pieghe dentro altre pieghe, all'infinito. In altre parole, un attrattore caotico è un frattale, cioè un oggetto che rivela particolari sempre più numerosi via via che viene ingrandito (*si veda l'illustrazione nella pagina a fronte*).

Il caos mescola le orbite nello spazio degli stati esattamente come un fornaio impasta il pane. Ci si può immaginare ciò che accade alle traiettorie vicine su un attrattore caotico versando nella pasta una goccia di colorante blu. L'impastatura è una combinazione di due azioni: lo stendimento della pasta, che fa diffondere il colorante, e il ripiegamento della pasta su se stessa. Dapprima la chiazza di colorante semplicemente si allunga, ma poi viene ripiegata e dopo un tempo piuttosto lungo si trova stirata e ripiegata molte volte. Osservandolo da vicino, si vede che l'impasto consiste in molti strati, alternativamente blu e bianchi. Già dopo 20 passaggi la lunghezza iniziale della macchia è aumentata più di un milione di volte e il suo spessore si è assottigliato fino a dimensioni molecolari. Il colorante blu è completamente mescolato con l'impasto. Il caos agisce allo stesso modo, ma invece di mescolare pasta mescola lo spazio degli stati. Ispirandosi a questa immagine del mescolamento, Otto E. Rössler, dell'Università di Tübingen, ha ideato l'esempio più semplice di attrattore caotico in un flusso (*si veda l'illustrazione a pagina 15*).

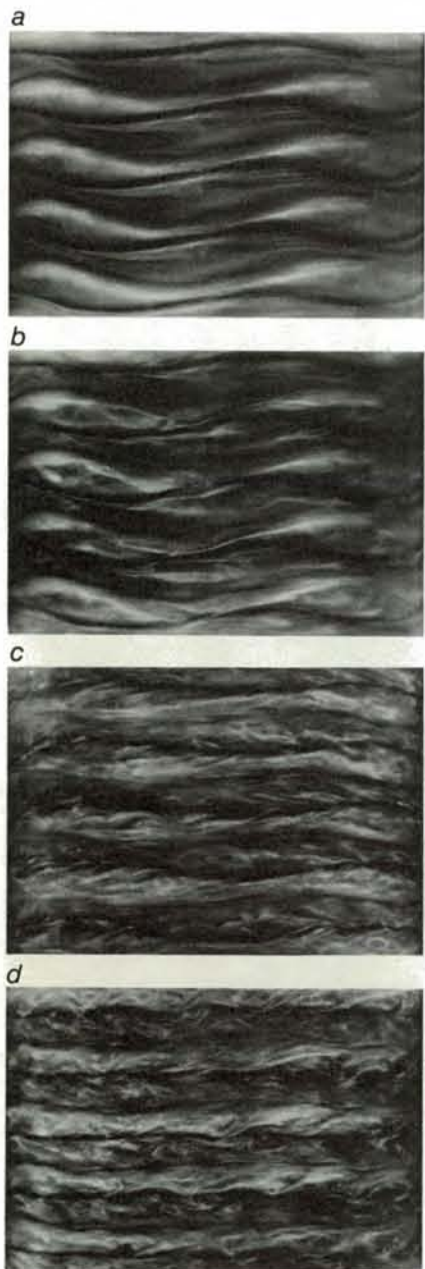
Quando si compiono osservazioni su un sistema fisico, è impossibile determinare esattamente lo stato del sistema a causa degli inevitabili errori di misurazione. Quindi lo stato del sistema non è situato in un unico punto, bensì all'interno di una piccola regione dello spazio degli stati. Benché sia l'indeterminazione quantistica che impone le dimensioni ultime di questa regione, in pratica vari tipi di rumore limitano la precisione delle misurazioni introducendo errori sostanzialmente più grandi. La piccola regione determinata da una misurazione corrisponde alla chiazza di colorante blu nell'impasto.

Quando si effettua una misurazione collocando il sistema in una piccola regione dello spazio degli stati, si ricava una certa quantità di informazione sul sistema. Quanto più accurata è la misurazione, tanto maggiore è la conoscenza che l'osservatore ricava sullo stato del sistema. Viceversa, quanto più ampia è la regione tanto più incerto è l'osservatore. Poiché nei sistemi non caotici i punti vicini rimangono vicini durante l'evoluzione temporale, una misurazione fornisce una certa quantità d'informazione che si conserva nel tempo. Questo è proprio il senso in cui questi sistemi sono prevedibili: le misurazioni iniziali contengono informazioni che possono essere sfruttate per prevederne il comportamento futuro. In altri termini, i sistemi dinamici prevedibili non sono molto sensibili agli errori di misurazione.

L'operazione di stiramento e piegatura di un attrattore caotico elimina sistematicamente l'informazione iniziale e la sostituisce con informazione nuova: lo stiramento amplifica le indeterminazioni su piccola scala, la piegatura avvicina traiettorie molto lontane tra loro e cancella l'informazione su grande scala. Quindi gli attrattori caotici si comportano come una sorta di pompa, poiché portano a manifestazione macroscopica le fluttuazioni microscopiche. È chiaro allora che non può esistere alcuna soluzione esatta, alcuna scorciatoia per prevedere il futuro. Dopo un breve intervallo di tempo l'indeterminazione corrispondente alla misura iniziale ricopre tutto l'attrattore e tutta la capacità di previsione è perduta: non vi è più alcun legame causale tra passato e futuro.

La divergenza delle traiettorie vicine è la ragione di fondo per la quale il caos porta all'imprevedibilità. Una misurazione perfetta corrisponderebbe a un punto dello spazio degli stati, ma tutte le misurazioni reali sono imprecise e generano un alone di indeterminazione. Lo stato effettivo può essere in qualunque punto interno a questo alone. Come si vede qui nel caso dell'attrattore di Lorenz, l'imprecisione della misura iniziale è rappresentata da 10 000 puntini rossi, che all'inizio sono così vicini tra loro da essere indistinguibili. Via via che ciascun punto si muove per effetto delle equazioni, l'alone viene stirato e forma un lungo filo sottile, che poi si piega su se stesso molte volte, fino a quando i punti risultano diffusi su tutto l'attrattore. Ora la previsione è impossibile: lo stato finale può essere in qualunque punto dell'attrattore. Per un attrattore prevedibile, invece, tutti gli stati finali rimangono vicini tra loro. I numeri sopra l'illustrazione sono in unità di 1/200 di secondo.

Gli attrattori caotici sono frattali, cioè oggetti che rivelano un numero sempre più grande di particolari via via che vengono ingranditi. Il caos genera frattali in modo naturale. Le traiettorie vicine si allontanano, ma perché il moto resti finito prima o poi devono ripiegarsi e riavvicinarsi tra loro. Ciò avviene ripetutamente e si generano così pieghe dentro le pieghe, all'infinito. Michel Hénon, dell'Osservatorio di Nizza, ha scoperto una regola semplice che stira e piega il piano, portando ogni punto in una posizione nuova. Da un unico punto iniziale sono stati tracciati tutti i punti successivi ottenuti applicando ripetutamente la regola di Hénon. La figura geometrica risultante (*a*) è un esempio di attrattore caotico. In *b* il quadratino rosso è ingrandito dieci volte. La ripetizione del procedimento (*c*, *d*) rivela nei particolari la struttura microscopica dell'attrattore. L'illustrazione in basso raffigura un bacino di attrazione della mappa di Hénon.

Gli attrattori caotici fungono localmente da amplificatori di rumore. Una piccola fluttuazione, dovuta magari a rumore termico, provoca dopo breve tempo un ampio scostamento nella posizione dell'orbita. Ma gli attrattori caotici differiscono dagli ordinari amplificatori di rumore per un aspetto importante. Poiché si suppone che lo stiramento e la piegatura siano ripetuti e continui, qualsiasi minuscola fluttuazione finisce per dominare il moto e il comportamento qualitativo è indipendente dal livello di rumore. Perciò non è possibile «calmare» direttamente i sistemi caotici, per esempio abbassandone la temperatura. I sistemi caotici generano aleatorietà di per sé, senza bisogno di alcun apporto aleatorio esterno. Il comportamento aleatorio scaturisce da qualche cosa di più di una semplice amplificazione degli errori e da una perdita della capacità di previsione; è dovuto alle orbite complesse generate da stiramento e piegatura.

Si osservi che tanto il comportamento caotico quanto quello non caotico si possono presentare in sistemi conservativi, cioè privi di dissipazione. Qui le orbite non tendono a disporsi su un attrattore, ma sono confinate su una superficie di energia. La dissipazione, però, è importante in molti sistemi reali, se non in tutti, ed è naturale aspettarsi che il concetto di attrattore abbia un'utilità generale.

Gli attrattori caotici con poche dimensioni hanno inaugurato un nuovo dominio della teoria dei sistemi dinamici, ma resta da stabilire se abbiano a che fare con l'aleatorietà dei sistemi fisici. La prima prova sperimentale a sostegno dell'ipotesi che alla base del moto aleatorio dei fluidi vi siano attrattori caotici fu piuttosto indiretta. L'esperimento fu condotto nel 1974 da Jerry P. Gollub dell'Haverford College e da Harry L. Swinney dell'Università del Texas ad Austin. La prova risultò indiretta perché i ricercatori si concentrarono non sull'attrattore in sé, ma sulle proprietà statistiche che lo caratterizzano.

Il sistema studiato era una cella di Couette, che è costituita da due cilindri concentrici; lo spazio fra i due cilindri è occupato da un fluido e uno o entrambi i cilindri vengono fatti ruotare a una certa velocità angolare. All'aumentare della velocità angolare il fluido manifesta configurazioni di flusso sempre più complesse, che dipendono dal tempo in modo complicato (*si veda l'illustrazione di questa pagina*). Gollub e Swinney in sostanza misuravano la velocità del fluido in un dato punto. All'aumentare della velocità di rotazione, osservarono prima una transizione da una velocità costante nel tempo a una velocità che variava periodicamente e infine a una velocità che variava in modo non periodico. Il cuore dell'esperimento era questa transizione al moto non periodico.

L'esperimento doveva distinguere tra due modelli teorici che prevedevano un diverso comportamento del fluido al variare della velocità di rotazione. Il modello di Landau per il moto aleatorio del fluido prevedeva che all'aumentare della velocità venisse eccitato un numero sempre maggiore di oscillazioni indipendenti del fluido. L'attrattore associato era un toro a molte dimensioni. Il modello di Landau era stato messo in discussione da David Ruelle dell'Institut des Hautes Études Scientifiques, presso Parigi, e da Floris Takens, dell'Università di Groningen, nei Paesi Bassi. Le loro argomentazioni matematiche inducevano a credere che non fosse plausibile che nel moto dei fluidi si presentasse l'attrattore associato con il modello di Landau. I risultati suggerivano invece che qualunque toro multidimensionale potesse dar luogo a un attrattore caotico, come aveva postulato per primo Lorenz.

Gollub e Swinney scoprirono che a bassa velocità di rotazione il flusso non variava nel tempo: l'attrattore soggiacente era un punto fisso. All'aumentare della velocità angolare l'acqua cominciava a oscillare con una sola frequenza indipendente e l'attrattore corrispondente era un ciclo limite (un'orbita periodica); quando poi la rotazione diveniva ancora più veloce, l'oscillazione assumeva due frequenze indipendenti, e l'attrattore corrispondente era un toro bidimensionale. La teoria di Landau prevedeva che, aumentando ulteriormente la velocità di rotazione, questa tendenza di comportamento venisse confermata, nel senso che dovessero via via comparire altre frequenze distinte. In corrispondenza di una velocità di rotazione critica, invece, compariva all'improvviso una gamma di frequenze continua. Quest'osservazione era in accordo con il «flusso non periodico deterministico» di Lorenz, e corroborava la sua idea che alla base della turbolenza dei fluidi vi siano gli attrattori caotici.

L'analisi di Gollub e Swinney avvalorava l'idea che gli attrattori caotici potessero essere alla base di alcuni moti aleatori dei fluidi, ma il loro lavoro non fu certo conclusivo. Sarebbe auspicabile dimostrare esplicitamente l'esistenza nei dati sperimentali di un attrattore caotico semplice. In genere un esperimento non registra tutti gli aspetti di un sistema, ma solo alcuni. Gollub e Swinney non riuscirono a registrare tutto il flusso della cella di Couette, ma solo la velocità del fluido in un dato punto. Il compito dello studioso è «ricostruire» l'attrattore dai dati limitati ottenuti. E ciò non è sempre possibile; se l'attrattore è troppo complicato qualcosa andrà perduto. In certi casi, però, è possibile ricostruire la dinamica a partire da dati limitati.

Una tecnica introdotta da noi e posta su solide basi matematiche da Takens consentì di ricostruire lo spazio degli stati e di cercare gli attrattori caotici. L'idea di base è che l'evoluzione di ciascuna componente di un sistema è determinata da quelle componenti con le quali interagisce. Pertanto nella storia di ciascuna componente sono contenute implicitamente informazioni sulle componenti con cui interagisce. Per ricostruire uno spazio degli stati «equivalente», si considera una singola componente e si trattano i valori misurati con ritardi temporali fissi (un secondo prima, due secondi prima, e così via, per esempio) come se fossero nuove dimensioni.

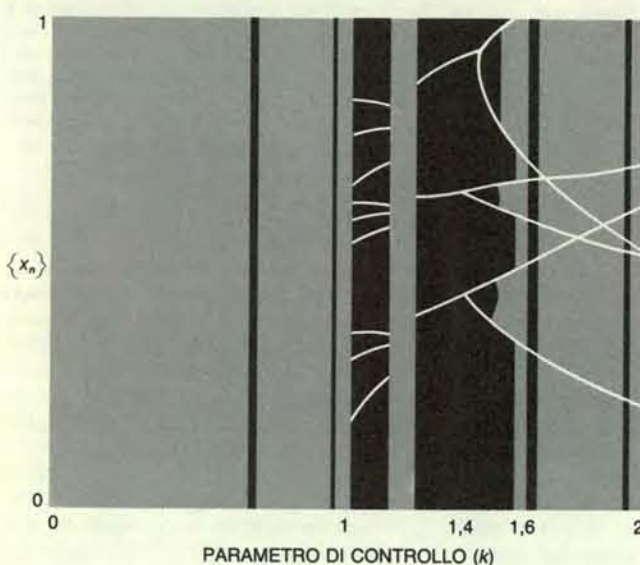
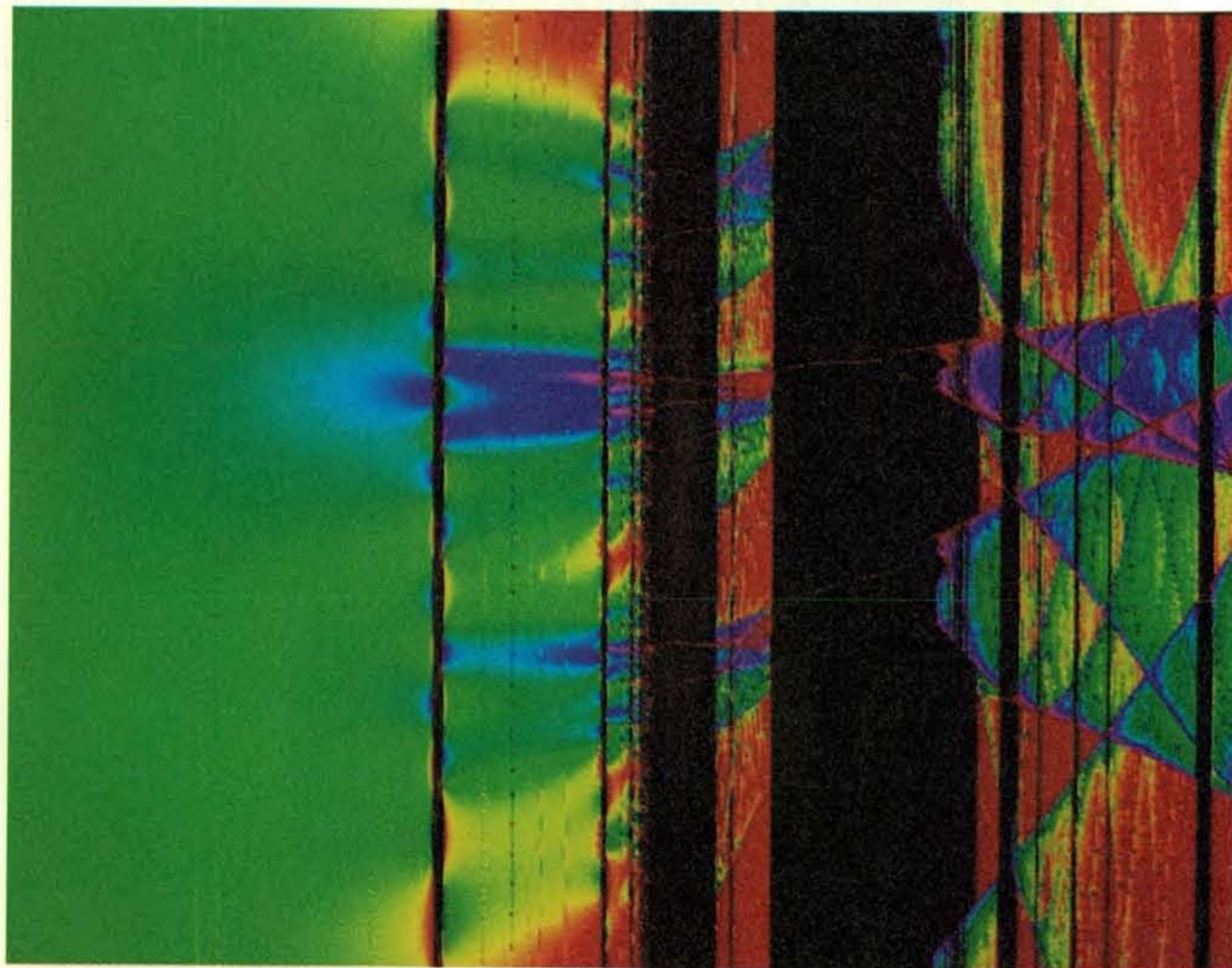
I risultati sperimentali confortano l'ipotesi che gli attrattori caotici siano alla base di alcuni tipi di moto aleatorio dei fluidi. L'acqua contenuta in una cella di Couette, che è costituita da due cilindri concentrici, è raffigurata in tre fotografie in successione. Lo spazio fra i due cilindri viene riempito d'acqua; quindi il cilindro interno viene posto in rotazione con una data velocità angolare (*a*). All'aumentare di questa velocità, il fluido manifesta un moto dall'andamento sempre più complesso (*b*): prima irregolare (*c*) e infine caotico (*d*).

Un rubinetto che gocciola è un esempio molto comune di sistema che può subire una transizione caotica. L'attrattore soggiacente viene ricostruito riportando a coppie in un grafico gli intervalli di tempo fra gocce successive, come appare nella parte in alto dell'illustrazione. Gli attrattori ricostruiti a partire da un reale rubinetto che gocciola (*a*, *c*) non sono di qualità inferiore agli attrattori generati mediante varianti della regola di Hénon (*b*, *d*). (L'attrattore di Hénon completo è a pagina 17.) Le illustrazioni *e* e *f* sono state ricostruite da flussi d'acqua ad alta velocità e rappresentano presumibilmente sezioni di attrattori caotici finora mai osservati. In ciascuno dei grafici sono state impiegate le coordinate ritardo temporale. La coordinata orizzontale è t_n, che è l'intervallo fra la goccia n e la goccia $n - 1$. La coordinata verticale è l'intervallo di tempo successivo, t_{n+1}, e la terza coordinata, ortogonale alla pagina, è t_{n+2}. Quindi ciascun punto è individuato da una terna di numeri (t_n, t_{n+1}, t_{n+2}) che sono stati tracciati per un insieme di 4094 campioni. Alle illustrazioni *b* e *d* è stato aggiunto rumore simulato.

I valori ritardati possono essere considerati come nuove coordinate, che definiscono un singolo punto in uno spazio degli stati multidimensionale. Ripetendo il procedimento e scegliendo i ritardi con riferimento a istanti diversi, si possono generare molti di questi punti. Si possono poi sfruttare altre tecniche per verificare se questi punti giacciono su un attrattore caotico o no. Benché questa rappresentazione sia per molti aspetti arbitraria, risulta che in essa vengono conservate le proprietà importanti di un attrattore, le quali non dipendono dai particolari della ricostruzione.

L'esempio che useremo per illustrare questa tecnica ha il pregio di essere noto e accessibile quasi a tutti. Molti sanno che le gocce che cadono da un rubinetto che gocciola presentano un andamento periodico. L'intervallo fra due gocce successive può essere regolarissimo, e molti, nelle notti d'insonnia, sono stati con l'orecchio teso ad aspettare che cadesse la goccia successiva. Meno noto è il comportamento di un rubinetto quando la velocità del flusso è un po' più elevata. Si può spesso individuare un regime in cui le gocce, pur cadendo sempre separate, cadono con un ritmo che non si ripete mai, come un tamburo che improvvisi cadenze sempre nuove. (È un esperimento che chiunque può effettuare facilmente; i rubinetti senza frangigetto funzionano meglio.) Il passaggio dal-

l'andamento ritmico a quello apparentemente aleatorio ricorda la transizione dal regime laminare a quello turbolento di un fluido. È possibile che alla base di questa aleatorietà vi sia un attrattore caotico semplice?

Lo studio sperimentale di un rubinetto che gocciola fu eseguito presso l'Università della California a Santa Cruz da uno degli autori (Shaw) in collaborazione con Peter L. Scott, Stephen C. Pope e Philip J. Martein. La prima variante dell'esperimento consisté nel far cadere le gocce da un normale rubinetto su un microfono e nel registrare i corrispondenti impulsi sonori. A pagina 19 sono illustrati alcuni risultati tipici di un esperimento un po' più raffinato. Un grafico in cui gli intervalli fra le gocce siano riportati a coppie è uno strumento efficace per ricavare una sezione dell'attrattore

La transizione al caos è illustrata da un diagramma di biforcazione, in cui una famiglia di attrattori (*asse verticale*) è rappresentata in funzione di un parametro di controllo (*asse orizzontale*). Il diagramma è stato generato da un semplice sistema dinamico che associa un numero a un altro. Il sistema dinamico impiegato è l'applicazione circolare ed è definito dall'equazione ricorsiva $x_{n+1} = \omega + x_n + k/2\pi \sin (2 \pi x_n)$. Per ciascun valore scelto del parametro di controllo k un calcolatore ha tracciato l'attrattore corrispondente. I colori esprimono la probabilità di trovare i punti sugli attrattori: il rosso corrisponde a regioni visitate spesso, il verde a regioni visitate meno spesso e il blu a regioni visitate di rado. Al crescere di k da 0 a 2 (*si veda il disegno a sinistra*), il diagramma illustra due cammini verso il caos: una strada quasiperiodica (da $k = 0$ a $k = 1$, che corrisponde alla regione verde qui sopra) e una strada di «raddoppio del periodo» (da $k = 1,4$ a $k = 2$). La strada quasiperiodica è equivalente, sotto il profilo matematico, a un cammino che passa attraverso un attrattore a forma di toro. Nella strada del raddoppio del periodo, che è basata sull'attrattore a ciclo limite, le diramazioni si presentano a coppie, secondo la serie geometrica 2, 4, 8, 16, 32 e così via. Le iterazioni oscillano tra le coppie dei rami. (Per un valore particolare di k - per esempio 1,4 - le iterazioni visitano soltanto tre valori. Per k più grande, questa «orbita di periodo tre» raddoppia il suo periodo e visita sei valori; poi, con un ulteriore raddoppio ne visita 12, e via di questo passo.) Da ultimo la struttura delle diramazioni diventa così fine che ne emerge una struttura a banda continua: si raggiunge una soglia oltre la quale fa la sua comparsa il caos.

soggiacente. Nel regime periodico, per esempio, il menisco dal quale si staccano le gocce si sposta in un modo continuo e ripartitivo che potrebbe essere rappresentato da un ciclo limite nello spazio degli stati. Ma nell'esperimento reale questo movimento continuo è inaccessibile; vengono registrati solo gli intervalli tra il distacco di una goccia e della successiva. È come quando si applica una luce stroboscopica a un moto regolare su una curva chiusa: se la cadenza è scelta bene, si vede solo un punto fisso.

Il risultato più interessante dell'esperimento fu che nel regime non periodico del rubinetto gocciolante furono effettivamente trovati attrattori caotici. L'aleatorietà delle gocce sarebbe potuta derivare da cause invisibili, per esempio da piccole vibrazioni o da correnti d'aria. Ma, se fosse stato così, non vi sarebbe stata alcuna relazione particolare fra un intervallo e il successivo e il grafico dei dati presi a coppie avrebbe avuto l'aspetto di una chiazza informe. Il fatto che nei grafici compaiano delle strutture dimostra che l'aleatorietà possiede una base deterministica. In particolare molti gruppi di dati hanno una forma a ferro di cavallo che è caratteristica del semplice processo di stiramento e piegatura descritto in precedenza. Questa forma caratteristica può essere considerata come l'«istantanea» di una piegatura in atto, per esempio una sezione parziale dell'attrattore di Rössler (*si veda l'illustrazione a pagina 15*). Altri gruppi di dati appaiono più complicati; potrebbero essere sezioni di attrattori a più dimensioni. Attualmente la geometria degli attrattori a più di tre dimensioni è quasi completamente sconosciuta.

Se un sistema è caotico, quanto è caotico? Una misura del caos è l'«entropia» del moto, che è grosso modo il tasso medio di stiramento e piegatura, ovvero il tasso medio al quale viene generata l'informazione. Un'altra stima è la «dimensione» dell'attrattore. Se un sistema è semplice, il suo comportamento dovrebbe essere descritto da un attrattore dotato di poche dimensioni nello spazio degli stati, come negli esempi forniti in questo articolo. Per individuare lo stato di un sistema più complicato potrebbe essere necessario usare parecchi numeri, e l'attrattore corrispondente avrebbe un numero maggiore di dimensioni.

La tecnica di ricostruzione, affiancata da misurazioni di entropia e di dimensione, consente di riesaminare il moto del fluido studiato da Gollub e Swinney. Ciò è stato fatto da alcuni ricercatori del gruppo di Swinney in collaborazione con due degli autori (Crutchfield e Farmer). La tecnica di ricostruzione ci ha permesso di ottenere immagini dell'attrattore soggiacente. Queste immagini non ci danno la sensazionale dimostrazione di un attrattore di poche dimensioni, come gli studi di altri sistemi, per esempio il rubinetto gocciolante. Tuttavia, le misurazioni dell'entropia e della dimensione rivelano che il moto irregolare del fluido in prossimità della transizione nella cella di Couette può essere descritto da attrattori caotici. All'aumentare della velocità di rotazione della cella, aumentano anche l'entropia e la dimensione degli attrattori soggiacenti.

Negli ultimi anni si è dimostrato che un numero crescente di sistemi ha un'aleatorietà dovuta a un attrattore caotico semplice. Fra questi il moto di convezione di un fluido riscaldato in una piccola scatola, i livelli oscillanti della concentrazione in una reazione chimica per rimescolamento, la contrazione delle cellule cardiache di pollo e un gran numero di oscillatori elettrici e meccanici. Inoltre si è dimostrato che i modelli al calcolatore di certi fenomeni, che vanno dalle epidemie all'attività elettrica di una cellula nervosa alle oscillazioni delle stelle, possiedono questo tipo semplice di aleatorietà. Sono in corso anche esperimenti diretti a cercare il caos in aree disparate quali le onde cerebrali e l'economia.

Si deve tuttavia sottolineare che la teoria del caos è ben lungi dall'essere una panacea. Molti gradi di libertà possono anche dar luogo a moti complicati che sono effettivamente aleatori. Anche quando si sappia che un dato sistema è caotico, questo fatto da solo non dice molto. Un buon esempio è costituito dagli urti reciproci e dai rimbalzi delle molecole di un gas. Sapere che questo sistema è caotico non rende più facile la previsione del suo comportamento. Le molecole del sistema sono così numerose che si può solo sperare in una descrizione statistica e le proprietà statistiche fondamentali possono essere ricavate senza prendere in considerazione il caos.

Vi sono altri problemi non ancora studiati, nei quali il ruolo del caos è ignoto. Che cosa si può dire per esempio di configurazioni spaziali continuamente mutevoli, come le dune del Sahara, o della turbolenza pienamente sviluppata? Non è chiaro se configurazioni spaziali complesse possano essere descritte utilmente da un unico attrattore in un unico spazio degli stati. Comunque l'esperienza con gli attrattori più semplici può forse contribuire a guidarci verso un quadro più articolato, che può comprendere interi complessi di forme deterministiche mobili nello spazio e affini agli attrattori caotici.

L'esistenza del caos ha conseguenze per il metodo scientifico stesso. Il metodo classico per verificare una teoria è fare previsioni e confrontarle poi con i dati sperimentali. Ma se i fenomeni sono caotici, le previsioni a lungo termine sono intrinsecamente impossibili e nel giudicare la bontà della teoria si deve tener conto anche di questo. Quindi il procedimento di verifica di una teoria diviene un'operazione molto più delicata, basata su proprietà statistiche e geometriche anziché su previsioni particolareggiate.

Il caos rappresenta una nuova sfida per il punto di vista riduzionistico, secondo cui un sistema può essere compreso scomponendolo e poi studiandone le singole parti. Questo punto di vista è stato largamente seguito nella scienza anche perché sono moltissimi i sistemi per i quali il comportamento del tutto è effettivamente la somma delle parti. Il caos, tuttavia, dimostra che un sistema può manifestare un comportamento complicato come risultato di un'interazione non lineare semplice tra poche componenti soltanto.

Il problema si sta facendo acuto in un'ampia gamma di discipline scientifiche, dalla descrizione della fisica microscopica alla costruzione di modelli per il comportamento macroscopico degli organismi biologici. Negli ultimi anni la capacità di ricavare conoscenze particolareggiate sulla struttura di un sistema ha compiuto progressi formidabili, ma la capacità di integrare queste conoscenze è stata ostacolata dall'assenza di una cornice concettuale adatta entro la quale descrivere il comportamento qualitativo. Anche quando si possegga una mappa completa del sistema nervoso di un organismo semplice, come il nematode studiato da Sidney Brenner dell'Università di Cambridge, non è possibile ricavarne il comportamento dell'organismo. Analogamente, la speranza che la fisica possa raggiungere la compiutezza grazie a una comprensione sempre più particolareggiata delle forze fisiche e dei costituenti fondamentali è infondata. L'interazione delle componenti a una data scala può provocare su scala più vasta un comportamento globale complesso che in generale non può essere ricavato dalla conoscenza delle singole componenti.

Spesso il caos è visto in termini delle limitazioni che comporta, come la mancanza di prevedibilità. Ma accade che la natura sfrutti il caos in modo costruttivo. Grazie all'amplificazione delle piccole fluttuazioni, esso può consentire ai sistemi naturali di accedere alla novità. Una preda che sfugga all'attacco di un predatore potrebbe usare un controllo caotico del volo come un elemento di sorpresa per sfuggire alla cattura. L'evoluzione biologica richiede la variabilità genetica, e il caos offre un mezzo per la strutturazione delle variazioni aleatorie, fornendo così la possibilità di porre la variabilità sotto il controllo dell'evoluzione.

Anche il progresso intellettuale è basato sull'introduzione di idee nuove e su nuove connessioni fra idee vecchie. La creatività innata potrebbe essere basata su un processo caotico, che amplifica selettivamente piccole fluttuazioni e le foggia in stati mentali macroscopici coerenti che vengono esperiti come pensieri. In certi casi i pensieri possono essere decisioni o essere percepiti come un esercizio della volontà. Sotto questa luce, il caos fornisce un meccanismo che spiega il libero arbitrio nell'ambito di un mondo retto da leggi deterministiche.

Il primo retrovirus umano

Causa una rara forma di leucemia e la sua scoperta, avvenuta nel 1978, ha posto le basi per poter identificare il virus affine, responsabile di una grave malattia, l'AIDS o sindrome da immunodeficienza acquisita

di Robert C. Gallo

In natura, il flusso dell'informazione genetica procede in generale dal DNA, sede dell'informazione genetica, all'RNA, che funge da tramite, e alle proteine, che sono le molecole funzionali della cellula. Per molti anni si è creduto che l'informazione genetica potesse in realtà fluire soltanto in una direzione e questa asserzione divenne nota come «dogma centrale» della biologia molecolare. Invece, non molti anni fa, il dogma centrale è stato sovvertito dalla scoperta di un gruppo di entità biologiche che capovolge l'ordine naturale delle cose. Il corredo genetico di queste entità, chiamate retrovirus, consiste di RNA. I retrovirus contengono anche un enzima, la transcriptasi inversa, che utilizza l'RNA virale come stampo per produrre il DNA; questo si integra nei cromosomi della cellula ospite e serve, quindi, come base per la replicazione virale.

La scoperta della transcriptasi inversa è stata stimolante non solo perché ha fatto luce su processi biologici fondamentali, ma anche perché ha contribuito a far capire in che modo i retrovirus sono causa di malattia. Era già noto che questi virus possono provocare negli animali il cancro, in prevalenza leucemie, cioè neoplasie dovute a trasformazione maligna dei globuli bianchi del sangue; sicché è parso assolutamente logico cercare analoghi virus cancerogeni negli esseri umani. Ma gli ostacoli, sia intellettuali sia tecnici, sono stati così rilevanti che solo nel 1978, quasi un decennio dopo la scoperta della transcriptasi inversa, i miei colleghi e io siamo riusciti a isolare il primo retrovirus umano.

Abbiamo chiamato l'oggetto da noi scoperto virus *T*-linfotropo, perché esso mostra un'attrazione (tropismo) per i linfociti *T*, globuli bianchi che hanno un ruolo di cruciale importanza nel modulare la risposta immunitaria. Il virus è cancerogeno ed entrando in un linfocito *T* può mettere in moto la catena di eventi che conduce alla leucemia. (Difatti, per questa ragione, viene talvolta chiamato virus della leucemia umana a linfociti *T*.) Utilizzando le tecniche dell'epidemiologia e della biologia molecolare abbiamo rapidamente scoperto parecchie cose su questo agente patogeno e sulla sua modalità d'azione. Ma, ciò che è più importante, abbiamo trovato altri virus che sono affini. Nel 1982, abbiamo scoperto un secondo retrovirus umano, che provoca anch'esso una leucemia. Così il virus umano *T*-linfotropo è diventato senza alcun indugio l'HTLV-I e il nuovo venuto è stato chiamato HTLV-II.

Ancora più sorprendente è stata la scoperta, nel 1983 e 1984, che l'agente responsabile della sindrome da immunodeficienza acquisita (AIDS) è un altro virus affine: l'HTLV-III. La sorpresa era in parte dovuta al fatto che AIDS e cancro hanno effetti opposti: mentre il cancro comporta una proliferazione incontrollata di linfociti *T*, l'AIDS li fa morire, paralizzando il sistema immunitario. Così, nell'arco di tempo di cinque anni dall'isolamento del primo retrovirus umano, sono state identificate due diverse categorie di questi virus, ciascuna con effetti patogeni diversi. Il modo in cui le due categorie sono state definite è una delle più affascinanti vicende della storia della biologia del XX secolo. Essa sarà esposta in due parti: la prima, cioè questo articolo, descrive la scoperta dell'HTLV-I, mentre la seconda, che costituirà l'argomento di un articolo del prossimo numero di «Le Scienze», racconterà la storia del virus responsabile dell'AIDS.

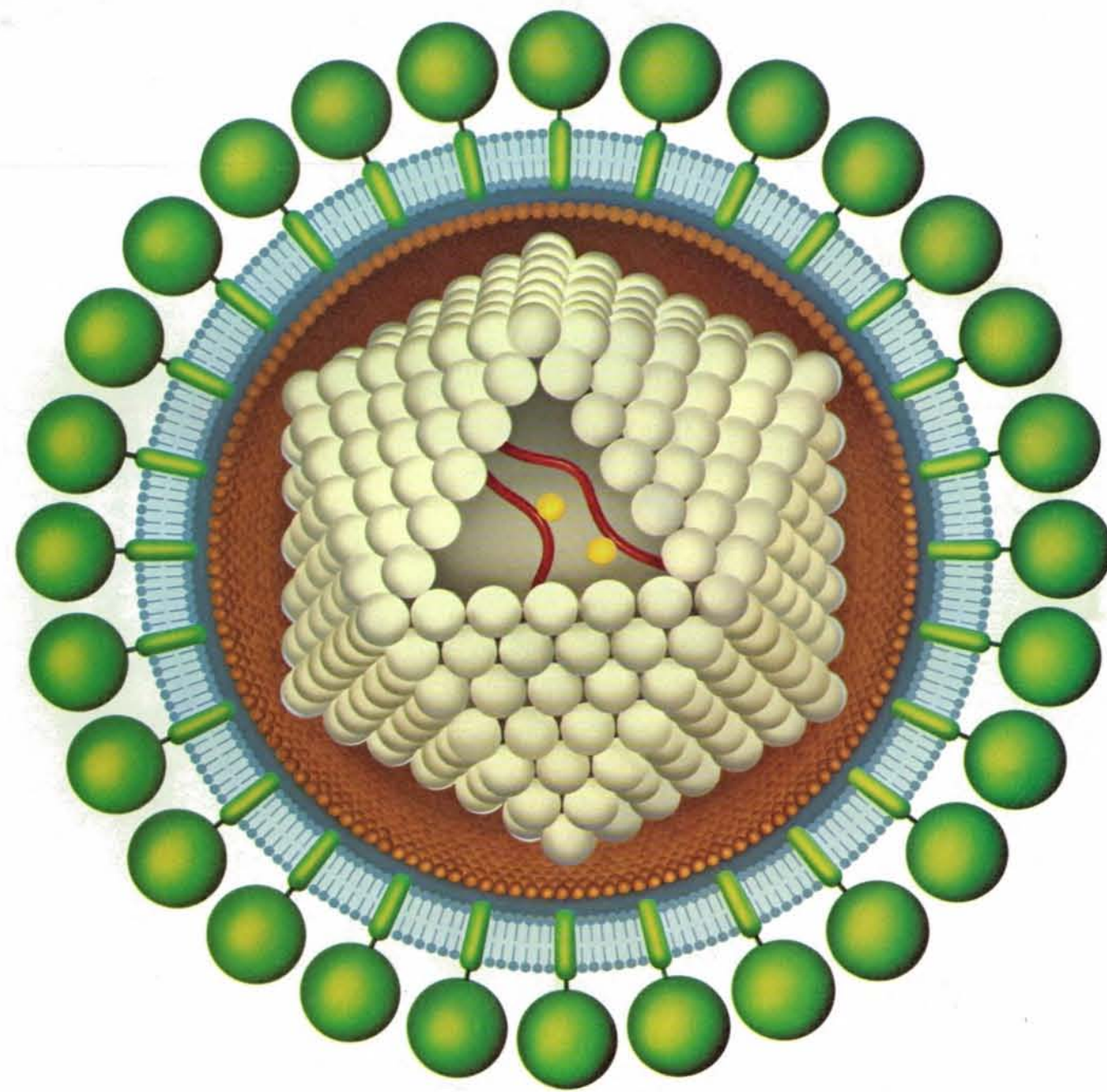
Lo scetticismo iniziale

Il prologo alla scoperta del primo retrovirus umano è una storia intrisa di scetticismo: in primo luogo, di dubbi sul fatto che agenti infettivi potessero risultare una causa importante di cancro e, in secondo luogo, dubbi sulla possibilità di individuare retrovirus negli esseri umani. La conferma che i retrovirus hanno effettivamente la capacità di provocare cancro, almeno in laboratorio, si è avuta piuttosto presto. Il primo retrovirus fu isolato nel 1910 da Peyton Rous del Rockefeller Institute for Medical Research, il quale dimostrò che quel virus, oggi chiamato virus del sarcoma aviario, può indurre neoplasie nei tessuti muscolare, osseo e vascolare dei polli.

I risultati di Rous incontrarono una così diffusa incredulità che Rous stesso decise di abbandonare la ricerca sui retrovirus; questo campo d'indagine conobbe pertanto un periodo di stasi totale, che durò fino agli anni cinquanta. In quel decennio e nel successivo, Ludwik Gross della Mount Sinai School of Medicine e altri ricercatori rilevarono che i retrovirus provocano neoplasie nei topi, nei polli e in altre specie. Ma la maggior parte dei biologi continuò ancora a dubitare che agenti infettivi avessero un ruolo significativo nella trasmissione del cancro al di fuori del laboratorio. Gli animali presi in considerazione nella ricerca sui retrovirus appartenevano, tutto sommato, a ceppi puri di laboratorio. Inoltre, molte infezioni erano congenite e sembrava improbabile che questa potesse essere una modalità di trasmissione importante negli esseri umani.

Pertanto, nel 1960, fu sostenuto, e in modo del tutto plausibile, che i retrovirus erano poco più di curiosità di laboratorio. Questo atteggiamento cambiò negli anni immediatamente successivi, quando William Jarrett dell'Università di Glasgow scoprì il virus della leucemia dei felini (FeLV). Jarrett, che aveva studiato veterinaria, dimostrò che quel virus era in grado di provocare non solo neoplasie delle cellule del sangue, ma anche aplasie (cioè, una crescita ridotta delle cellule colpite) e una deficienza immunitaria analoga a quella che, in seguito, sarebbe stata osservata nei soggetti colpiti da AIDS. Fatto altrettanto importante, Jarrett (con il proprio fratello e collega Oswald Jarrett e con Myron Essex della Harvard University e William D. Hardy, Jr., del Memorial Sloan-Kettering Cancer Center) dimostrò anche che le malattie provocate dal FeLV si trasmettevano tra gatti non affini viventi nel medesimo ambiente domestico.

Con la scoperta del FeLV, i retrovirus superarono lo stato di pura curiosità di laboratorio. Malgrado questo, non si notò alcuna fretta di isolare analoghi agenti infettivi negli esseri umani. Il motivo è da ricercare nel fatto che, allora, l'attenzione era attratta dai retrovirus endogeni. Questi sono virus il cui materiale genetico si riscontra talvolta nei cromosomi di molte specie animali, presumibilmente come residuo di una vecchia infezione. In alcuni casi, tali sequenze endogene di DNA possono dare origine a particelle virali che siano funzionalmente attive e infettive. Nell'insieme, tuttavia, esse sono trasmesse ad altri membri della stessa specie attraverso la normale eredità mendeliana, cioè attraverso le cellule uovo e gli spermatozoi. Nel 1969, Robert J. Huebner e George J. Todaro del National Cancer Institute proposero che l'attivazione di queste sequenze endogene, normalmente silenti, da parte di agenti cancerogeni fosse il meccanismo alla base di ogni forma di neoplasia.

Da allora, per ironia della sorte, tutte le malattie note per essere provocate in natura da retrovirus sono risultate causate da virus esogeni e i virus endogeni sono normalmente considerati come un enigma dell'evoluzione, con scarsa pertinenza clinica. Alla fine degli anni sessanta, tuttavia, i virus esogeni, nell'ecci-

Questa rappresentazione di un virus mette in rilievo buona parte delle conoscenze attuali sulla struttura del primo retrovirus umano che è stato identificato: l'HTLV-I o virus umano *T*-linfotropo di tipo I. Essa mostra il virione, o particella virale, in sezione. La particella ha un diametro di circa 100 nanometri (un decimillesimo di millimetro). Il suo involucro esterno è costituito da un duplice strato di molecole lipidiche intercalate a molecole proteiche (*in verde*) e ricopre una porzione centrale che contiene svariati tipi di proteine (*in marrone chiaro e in avorio*). In questa parte centrale vi sono anche due molecole di RNA (*in rosso*) nel quale si trova registrata l'informazione genetica necessaria al virus per sintetizzare i propri componenti e quindi per riprodursi. Associate all'RNA vi sono inoltre diverse copie di transcriptasi inversa (*in giallo*), un enzima che sfrutta l'RNA virale come stampo per il montaggio di una corrispondente molecola di DNA a duplice filamento.

1

2 GLICOPROTIDE DELL'INVOLUCRO

PROTEINA DELLA PORZIONE CENTRALE

MEMBRANA CELLULARE

3

RNA

TRANSCRIPTASI INVERSA

4

RNA

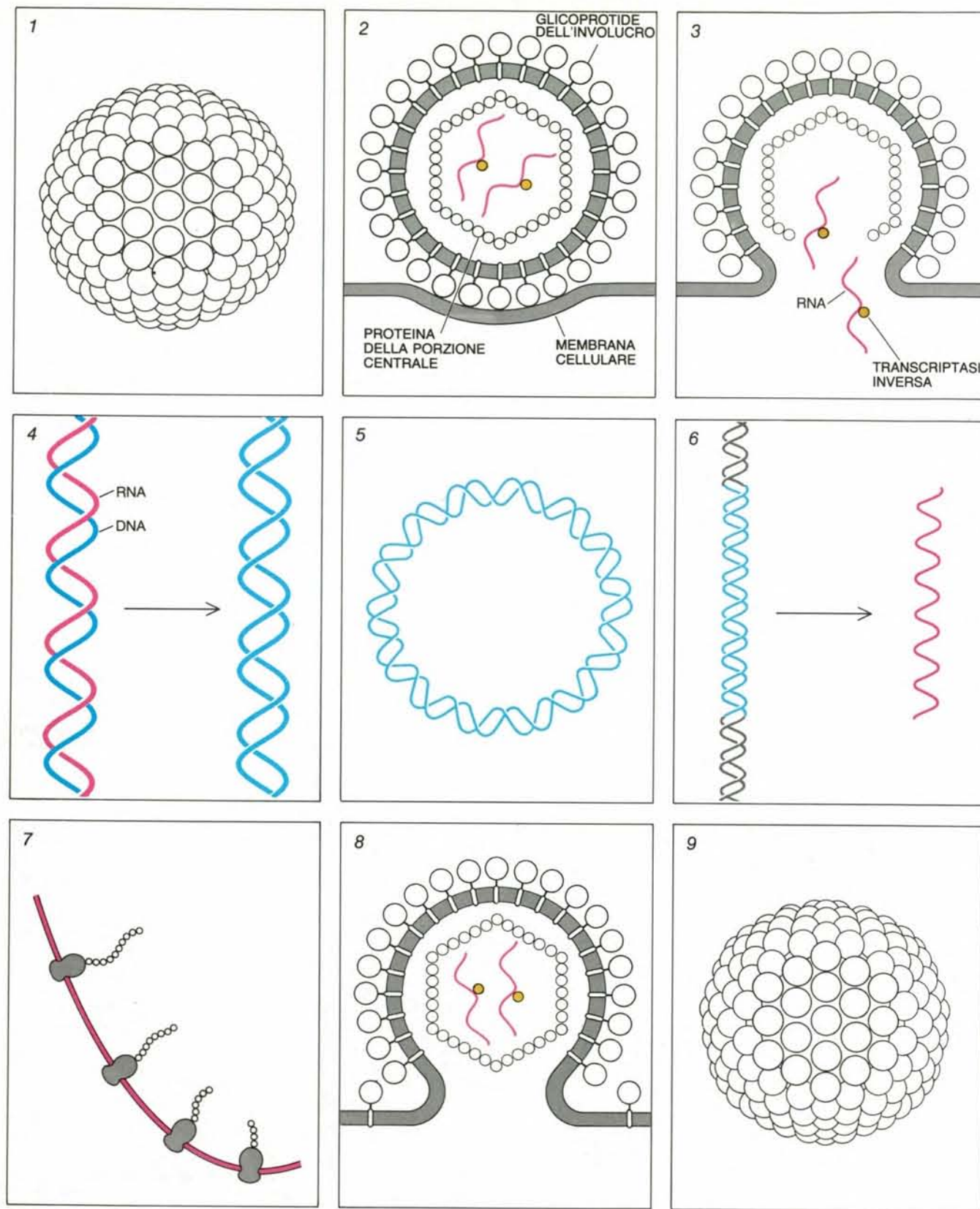
DNA

5

6

7

8

9

Il ciclo vitale di un retrovirus comprende l'integrazione del genoma virale (il corredo genetico completo) nel DNA della cellula ospite. La particella virale (*1*) interagisce con la membrana della cellula ospite (*2*), la quale nel caso dell'HTLV-I è un globulo bianco chiamato linfocito *T*. L'involucro del virus si fonde con la membrana esterna della cellula (*3*) in modo che il contenuto del virione si riversi nel citoplasma. Qui, la transcriptasi inversa produce un filamento singolo di DNA, che corrisponde all'RNA virale; successivamente (allorché l'RNA si degrada), viene prodotto un secondo filamento di DNA (*4*). Il DNA a duplice filamento migra verso il nucleo della cellula ospite, dove forma una struttura circolare (*5*), e si inserisce a caso in un cromosoma dell'ospite (*6*). In seguito, il DNA virale può essere trascritto in RNA, che a sua volta viene tradotto in proteina sui ribosomi presenti nel citoplasma (*7*). Le proteine neosintetizzate e l'RNA si riuniscono e formano una protuberanza all'esterno (*8*), dando origine a un nuovo virione che incorpora materiale lipidico ricavato dalla membrana esterna della cellula (*9*).

tazione generale provocata dai loro affini endogeni, che apparivano un poco dovunque, non venivano presi granché in considerazione.

La trascrizione inversa

Il favore incontrato dalla teoria di Huebner-Todaro non fu l'unico problema che, nel 1969, dovettero affrontare coloro che erano interessati ai virus esogeni. Esisteva anche una difficoltà concettuale riguardo al modo in cui il genoma del virus, che consiste notoriamente di RNA, interagiva con i geni della cellula ospite inducendo una neoplasia. Howard M. Temin della Medical School dell'Università del Wisconsin aveva suggerito che il ciclo vitale di un retrovirus esogeno includesse una fase intermedia a DNA, il provirus, ma i particolari di questo fatto non risultavano chiari.

La scoperta della transcriptasi inversa fatta da Temin (un risultato raggiunto indipendentemente e simultaneamente da David Baltimore del Massachusetts Institute of Technology) offrì una base concreta alla teoria del provirus e i particolari del ciclo vitale trovarono subito la loro esatta collocazione (si veda l'articolo *Sintesi di DNA diretta dall'RNA* di Howard M. Temin in «Le Scienze», n. 44, aprile 1972). Fu chiaro allora che, quando un retrovirus infetta una cellula, la sua transcriptasi inversa sintetizza una molecola di DNA che corrisponde al codice contenuto nell'RNA virale. Quel DNA può poi farsi strada verso il nucleo della cellula, dove si inserisce, integrandosi, in mezzo ai geni dell'ospite. In seguito, per interazione con questi geni, il provirus può dar luogo a un tumore. Poiché il DNA virale porta l'informazione necessaria a sintetizzare i componenti del virus, può anche essere attivato in modo da formare nuove particelle virali.

Nel 1970, quando Temin e Baltimore scoprirono la transcriptasi inversa, stavo studiando le DNA-polimerasi nelle cellule del sangue. Le DNA-polimerasi sono enzimi che servono a «costruire» le molecole di DNA e la transcriptasi inversa è un membro del loro gruppo, anche se un membro insolito. Sotto l'influenza delle idee di Temin, decisi di cercare la transcriptasi inversa nelle cellule leucemiche umane, sperando di trovare in esse un retrovirus.

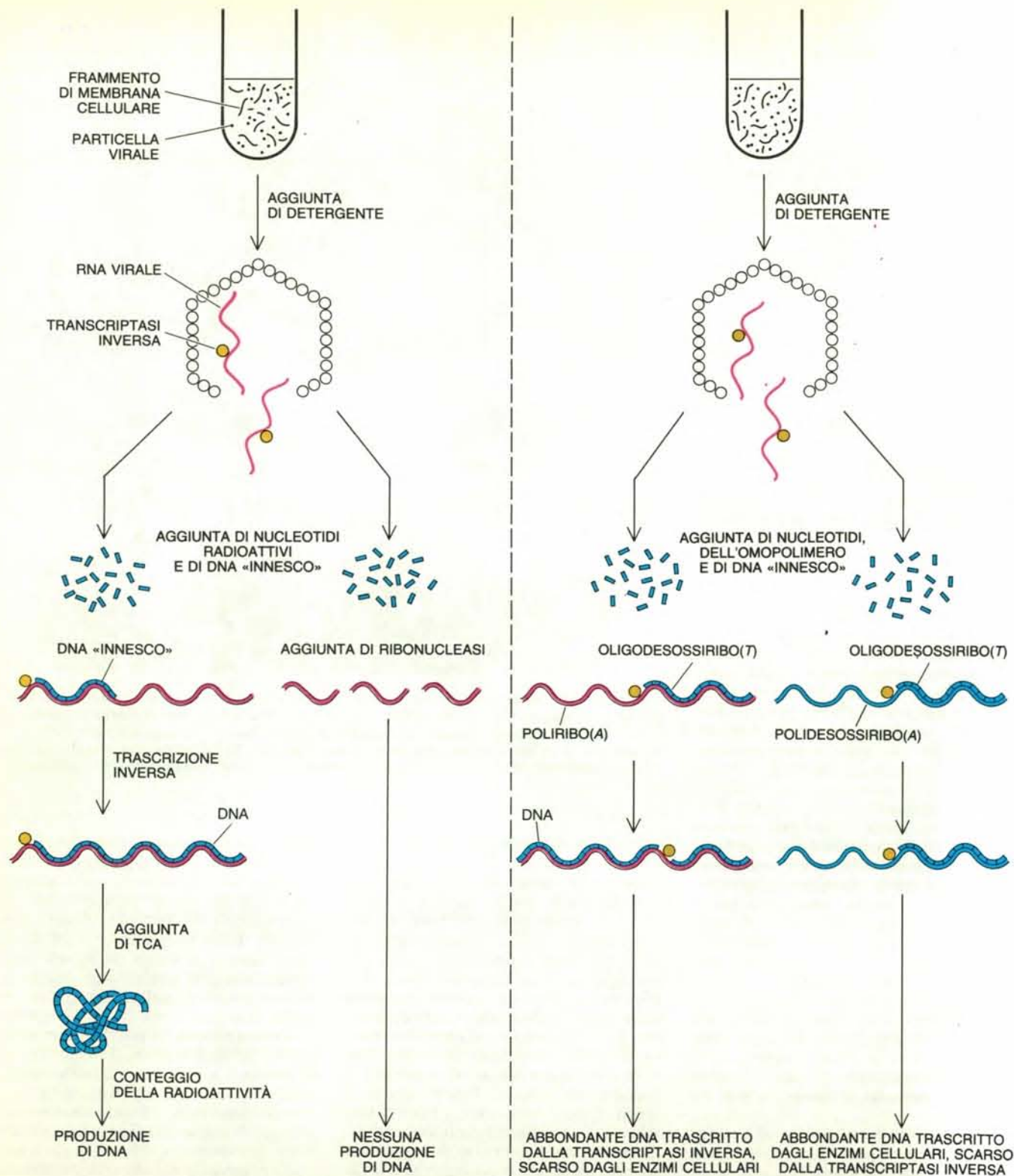
Nel far questo andavo contro il buonsenso. I virus responsabili della leucemia negli animali (già allora ben conosciuti nei polli, nei topi e nei gatti) vanno incontro a una diffusa replicazione prima che compaia la malattia e le nuove particelle virali vengono facilmente evidenziate al microscopio elettronico. Accurate osservazioni di cellule leucemiche umane al microscopio elettronico non avevano mai fornito immagini di questo tipo e avevano fatto concludere alla maggior parte dei ricercatori che non esistessero retrovirus nella specie umana.

Eppure a me - e, indipendentemente, anche a Sol Spiegelman della Columbia University - pareva che la ricerca non fosse priva di senso. Pensavo che forse i retrovirus della specie umana avrebbero potuto sfruttare un diverso meccanismo cancerogeno, che non comportasse affatto un'estesa replicazione virale. In questo caso, il microscopio elettronico (uno strumento poco efficace per questo scopo) non avrebbe mai rivelato l'agente patogeno. Un saggio più sensibile, invece, avrebbe potuto farlo: fortunatamente ce n'era uno a portata di mano per la transcriptasi inversa. Poiché questa è specifica per i retrovirus, il fatto di trovarla nelle cellule tumorali avrebbe indicato la presenza in esse di quei virus. Inoltre, un saggio biochimico atto a valutare la capacità dell'enzima di costruire la molecola del DNA poteva essere reso di gran lunga più specifico e più sensibile della microscopia.

Nuove particelle virali sporgono come gemme da un linfocito *T* in coltura: esse appaiono come piccoli cerchietti scuri alla periferia della cellula. La struttura irregolare al centro di questa è il nucleo; le aree rotondeggianti attorno sono goccioline lipidiche. L'immagine è rara: l'HTLV-I si replica solo in un breve periodo del ciclo vitale del linfocito *T*. La microfotografia elettronica, con ingrandimento di 13 000 diametri, è di Bernhard Kramarsky della Electro-Nucleonics, Inc.

I pregi delle colture in vitro

Tra il 1970 e il 1975 il gruppo di Spiegelman e il mio si sono dati da fare per perfezionare il saggio per la transcriptasi inversa, fino a fargli raggiungere una sensibilità di parecchi ordini di grandezza superiore a quella della microscopia elettronica (*si veda l'illustrazione di pagina 26*). Ricorrendo a prove del genere, i miei colleghi e io abbiamo ottenuto alcuni risultati allettanti, ma ancora non decisivi. Dalle cellule leucemiche di alcuni pazienti abbiamo purificato DNA-polimerasi che sembravano dotate di tutte le proprietà della transcriptasi inversa. Si notava, tuttavia, un'ambiguità: gli enzimi sarebbero potuti essere polimerasi cellulari insolite, riconoscibili solo perché il loro numero aumenta nelle cellule malate. Per dimostrare la loro origine in modo inequivocabile avevamo bisogno di una scorta di gran lunga maggiore dell'ipotetico virus, il quale a sua volta esigeva da noi che coltivassimo le cellule che infettava.

Ora, le cellule del sangue umano, sane o neoplastiche, sono difficilmente coltivabili in laboratorio e, in effetti, all'inizio degli anni settanta non esisteva alcuna linea cellulare in laboratorio che fosse adatta ai nostri scopi. A quell'epoca, però, erano state scoperte alcune proteine, chiamate fattori di crescita, in grado di stimolare la crescita in laboratorio di altre cellule restie a farlo. Sembrava dunque ragionevole pensare che quelle pro-

Il saggio per la transcriptasi inversa si basa sull'identificazione della capacità di questo enzima di costruire una molecola di DNA a partire da uno stampo di RNA. Nel protocollo originario (*a sinistra*), veniva utilizzato come stampo RNA virale endogeno. Le cellule infettate venivano centrifugate per separare le particelle virali (assieme a frammenti di membrana cellulare) da altro materiale. Un detergente blando era aggiunto poi per demolire le particelle, mettendo in libertà la transcriptasi inversa. Le subunità del DNA (nucleotidi), marcate con elementi radioattivi, e un frammento specifico di DNA, che fungeva da «innesco» per il montaggio del DNA, venivano quindi aggiunti alla soluzione contenente l'enzima. La transcriptasi inversa sintetizzava il DNA, che veniva successivamente fatto precipitare mediante acido tricloroacetico (TCA). La misurazione della radioattività del precipitato ha mostrato che i nucleotidi marcati erano stati effettivamente incorporati nel DNA. Se alla soluzione veniva aggiunta la ribonucleasi (un enzima che demolisce l'RNA), non si aveva la formazione di DNA, il che dimostra che il processo di formazione del DNA dipende da uno stampo di RNA. In seguito, il saggio è stato modificato (*a destra*) utilizzando stampi sintetici, il che ha permesso di distinguere in maniera chiara e indubitabile tra l'attività della transcriptasi inversa e l'attività di enzimi cellulari affini che utilizzano il DNA come stampo. Se viene utilizzato lo stampo chiamato poliribo(*A*), che è una forma di RNA, l'enzima virale produce molto DNA mentre gli enzimi cellulari ne producono poco. Se, invece, si usa lo stampo polidesossiribo(*A*), una forma di DNA, la situazione si capovolge.

teine potessero essere la chiave per coltivare i globuli bianchi del sangue. Di conseguenza, fu avviata una ricerca per trovare i fattori di crescita che erano adatti ai nostri scopi.

Un passo in questa direzione era stato compiuto nel corso degli anni sessanta da Peter C. Nowell della School of Medicine dell'Università della Pennsylvania. Nowell aveva trovato che una proteina, derivata dalle piante e chiamata fitoemagglutinina (PHA), poteva indurre certi globuli bianchi a crescere, a diventare attivi e a dividersi uno o due volte. In seguito, venne mostrato che le cellule di Nowell erano linfociti *T* e l'esperimento che questo ricercatore aveva effettuato era in pratica una coltivazione *in vitro* a breve termine. Nel 1976, riprendendo il lavoro da dove Nowell l'aveva lasciato, i miei collaboratori Doris Morgan e Francis Ruscetti, assieme a me, hanno trovato che alcuni linfociti *T*, dopo essere stati stimolati con la fitoemagglutinina, liberano effettivamente un fattore di crescita. Abbiamo chiamato questa nuova sostanza fattore di crescita dei linfociti *T*, o TCGF; oggi essa viene chiamata in generale interleuchina-2, o semplicemente IL-2.

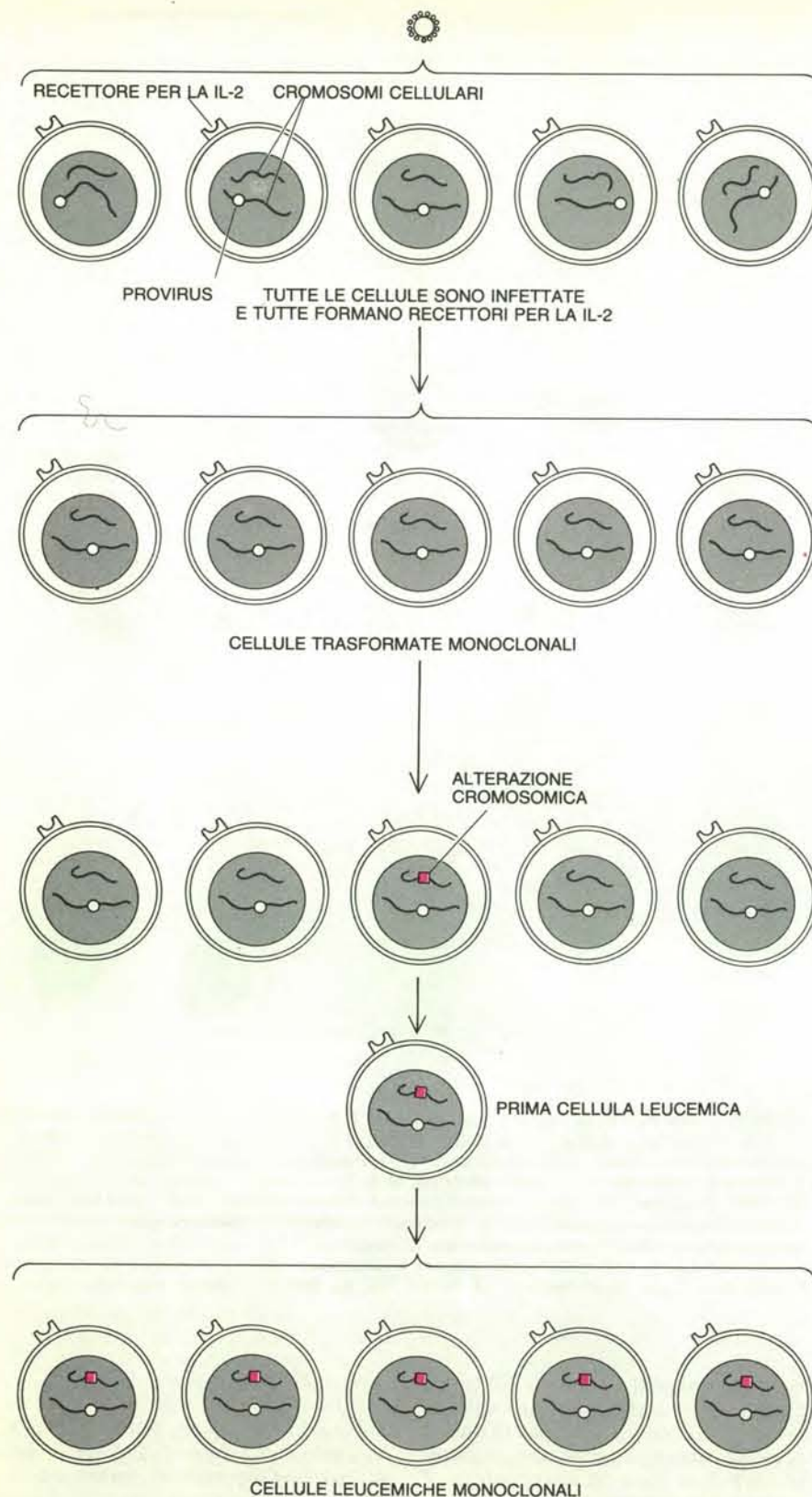
I linfociti *T* attivati non solo secernono la IL-2, ma formano anche molecole alla loro superficie, che fungono da recettori per quella proteina. Quando l'interleuchina-2 si lega al recettore, le cellule cominciano a dividersi. Avevamo dunque trovato un modo per coltivare i linfociti *T*: dopo aver accumulato una sufficiente quantità di IL-2, potevamo aggiungerla alle cellule prestimolate e quindi mantenerne la crescita per lunghi periodi. Per combinazione, questa sequenza di eventi è analoga a quella che si svolge normalmente nel sistema immunitario. I linfociti *T*, che derivano da un precursore nel midollo osseo, migrano nel timo per diventare maturi. In questa sede si differenziano in due classi: *T*4 e *T*8, con differenti molecole marcatrici sulla loro superficie e con differenti funzioni immunitarie. Come reazione all'infezione, alcuni di questi linfociti *T* maturi vengono dapprima attivati dalla proteina di nome IL-1 e quindi stimolati a dividersi da IL-2.

Alla fine degli anni settanta, queste vie erano ancora percepite in modo indistinto. Fu solo più tardi che le ricerche degli immunologi, in particolare di Kendall A. Smith e collaboratori della Dartmouth Medical School e di Hans Wigzell del Karolinska Institutet di Stoccolma, misero in luce il pieno significato di IL-2 per il sistema immunitario. In ogni caso, il nostro interesse primario non era indirizzato alla risposta immunitaria come tale, ma alla coltivazione dei linfociti *T* e, per questo scopo, la IL-2 risultò molto efficace.

Difatti Bernard Poiesz, un ricercatore che aveva continuato a lavorare nel nostro gruppo dopo il dottorato, trovò che alcuni linfociti *T* leucemici umani possono essere coltivati con la IL-2 senza una precedente attivazione da parte della fitoemagglutinina. La funzione di questa proteina consiste nel generare i recettori per la IL-2, di cui sono privi i linfociti *T* normali. Chiaramente i linfociti *T* leucemici possedevano questi recettori e quindi non avevano più bisogno di attivazione. Come dirò più avanti, questo sorprendente risultato ha un evidente significato per il modo in cui l'HTLV-I provoca il cancro. Ulteriori ricerche sui linfociti *T* leucemici hanno rivelato che essi erano relativamente maturi e possedevano la molecola marcatrice *T*4.

I linfociti *T* sono attivati in risposta a un'infezione. Il processo parte da un globulo bianco, il macrofago, che incontra un antigene, proteina di superficie di un organismo invasore. Il macrofago secerne una proteina, la interleuchina-1 (IL-1). Quando sono raggiunti da questa, i linfociti *T* quiescenti secernono una seconda proteina, la IL-2, e formano recettori per essa alla loro superficie. Il legame della IL-2 con i propri recettori induce il linfocito *T* a dividersi e ad andare incontro a maturazione (assumendo le caratteristiche funzionali necessarie per svolgere il suo ruolo complesso nella risposta immunitaria). La scoperta che l'IL-2 (in origine chiamato fattore di crescita del linfocito *T* o TCGF) poteva essere utilizzato per coltivare i linfociti *T* in laboratorio è stata una tappa fondamentale nel lavoro che ha portato l'autore a isolare l'HTLV-I.

I primi HTLV-I isolati

Proprio dai linfociti *T* coltivati con la IL-2 abbiamo isolato nel 1978-1979 i primi HTLV-I. Quei linfociti appartenevano a due pazienti affetti da leucemia. I miei collaboratori e io abbiamo ottenuto il virus, lo abbiamo caratterizzato e abbiamo dimostrato che era un virus specificamente umano. I nostri risultati sono stati poi pubblicati nel 1980 e all'inizio del 1981. In seguito abbiamo isolato molti altri esempi di HTLV-I e Marvin Reitz, del nostro gruppo, ha dimostrato che tutti questi virus non erano affatto

L'HTLV-I provoca la leucemia mediante un processo a più fasi. Quando il virus infetta un gruppo di linfociti *T*, questi producono recettori per la IL-2 e in ciascuno di essi si trova il provirus (cioè il DNA virale integrato nel genoma del linfocito). In ogni linfocito, però, il provirus si trova in una posizione differente. Alcuni linfociti infettati sono trasformati e passano a uno stato precanceroso; una parte di essi dà quindi origine a cloni di discendenti, ciascuno dei quali contiene una copia identica del DNA del progenitore, che include il provirus. In una cellula di un clone di questo tipo un'ulteriore alterazione cromosomica può condurre alla definitiva comparsa di uno stato canceroso. Tutte le cellule leucemiche derivano da questo singolo antenato canceroso.

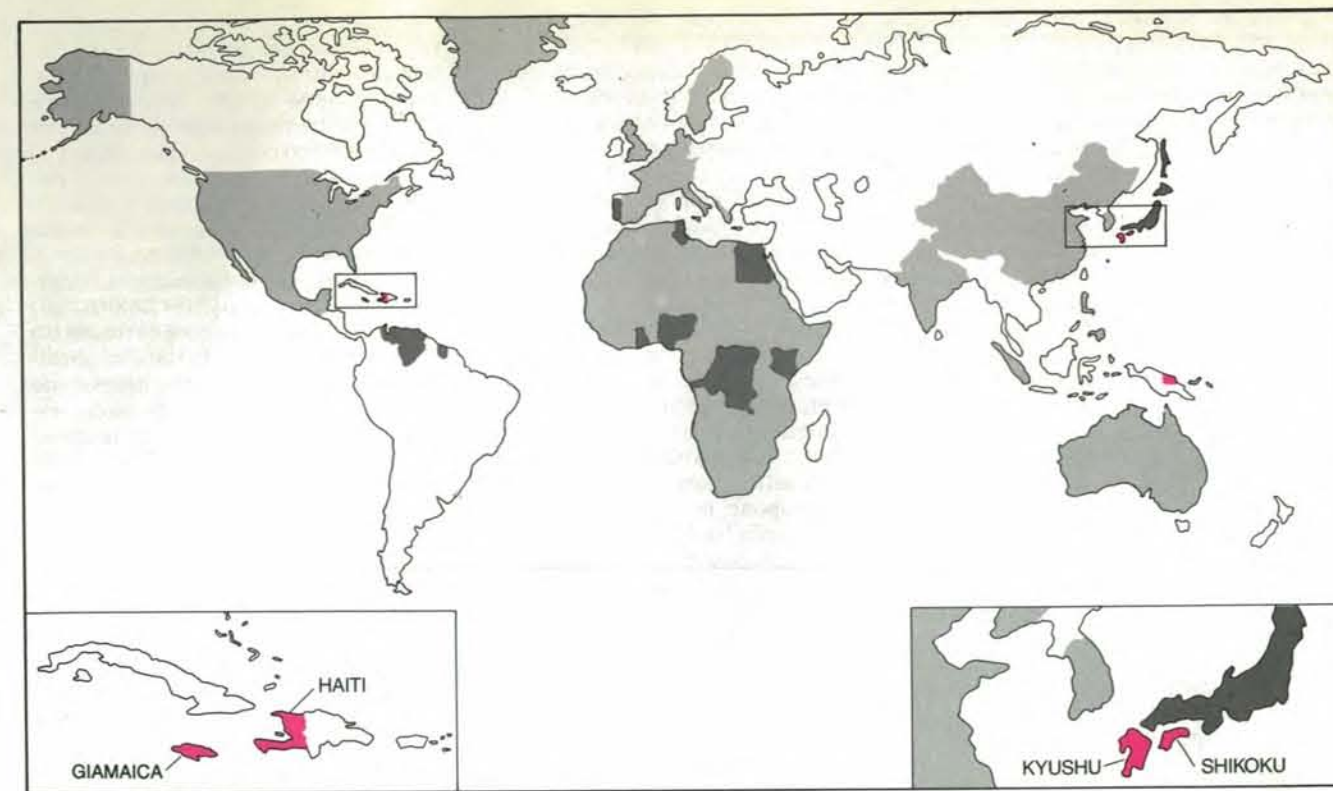
legati da stretta affinità ai virus animali precedentemente descritti. Fatto ugualmente importante, i virus da noi isolati non erano endogeni, bensì esogeni; pertanto, essi erano dello stesso tipo di virus che provoca malattia negli animali.

Le idee erano però piuttosto confuse sul tipo di malattia alla quale era associato l'HTLV-I. E la questione doveva proprio essere questa, visto che l'origine della leucemia a linfociti *T* non era ben conosciuta e che, di conseguenza, i clinici erano stati costretti a basarsi sui sintomi per suddividere in categorie le differenti forme di leucemia. I pazienti dai quali abbiamo isolato per la prima volta l'HTLV-I presentavano neoplasie a carico dei linfociti *T4* maturi, associate ad anomalie cutanee risultanti da infiltrazione della cute da parte di cellule del sangue neoplastiche.

Un simile quadro clinico è riferibile alla micosi fungoide o alla leucemia a linfociti *T* di Sézary. Era, tuttavia, chiaro dai nostri precedenti studi che l'HTLV-I poteva essere individuato soltanto in una piccola frazione di pazienti affetti dalle suddette sindromi. Per contro, molti pazienti leucemici che davano una reazione positiva per l'HTLV-I non mostravano anomalie cutanee e pertanto non rispondevano alla definizione clinica della micosi fungoide o della sindrome di Sézary.

Un poco di luce venne fatta in questa oscurità quando si venne a conoscenza di una malattia chiamata leucemia a linfociti *T* dell'adulto (ATL), descritta per la prima volta nel 1977-1978 da Kiyoshi Takatsuki dell'Università di Kyoto. L'ATL assomiglia alla micosi fungoide e alla sindrome di Sézary, ma si notano tra essa e queste sindromi alcune differenze significative. Tanto per cominciare la cute non è sempre interessata. Inoltre, l'ATL colpisce in modo più grave delle altre due malattie (la sopravvivenza media dal momento della diagnosi è di solo tre o quattro mesi). Le sue vittime, benché adulti, sono di solito più giovani e i linfociti *T* neoplastici, presenti nel sangue, sono spesso associati a una ipercalcemia (aumento del contenuto di calcio nel sangue). L'ipercalcemia da sola può essere fatale, ma più frequentemente la morte consegue a una proliferazione esplosiva di cellule leucemiche o a infezioni opportuniste.

Una geografia emergente

Takatsuki e collaboratori hanno notato che l'ATL presenta una concentrazione molto elevata nelle isole di Kyushu e Shikoku, le più meridionali tra le maggiori isole del Giappone: ciò ha fatto pensare che la malattia potesse essere causata da un agente infettivo. Poiché i sintomi ricordavano quelli dei casi che avevano fornito l'HTLV-I, abbiamo deciso di collaborare con Yohei Ito dell'Università di Kyoto per scoprire se i pazienti affetti da ATL ospitassero anch'essi

La distribuzione a livello mondiale dell'infezione da HTLV-I fornisce indizi sull'origine del virus. Le principali zone endemiche del virus (*in colore*) sono le regioni caraibiche e il Giappone sudoccidentale. Un'altra possibile zona endemica, finora non confermata, sarebbe la Nuova Guinea. Zone con livelli di infezione intermedi (*in grigio intenso*) includono parti dell'Africa equatoriale e dell'America latina. In altre zone si riscontrano bassi livelli di infezione (*in grigio chiaro*). Per buona parte del mondo (*in bianco*), i dati epidemiologici sono, tuttavia, scarsi. La scoperta di virus molto affini all'HTLV-I in certe specie di scimmie africane induce a supporre che il virus abbia avuto origine in Africa e poi si sia diffuso nel resto del mondo con il commercio, in particolare mediante il mercato degli schiavi. I dati riportati dalla mappa sono stati raccolti da William A. Blattner in collaborazione con Carl Saxinger, Marjorie Robert-Guroff e l'autore (tutti del National Cancer Institute).

il virus. E in effetti è stato così. Il siero di tutti i pazienti affetti da ATL, analizzato, ha mostrato di contenere anticorpi che reagivano con l'HTLV-I. Ma, ciò che è più importante, gli anticorpi monoclonali (anticorpi specifici che reagiscono con una proteina soltanto) prodotti contro le proteine dell'HTLV-I reagivano con le cellule colpite da ATL, ma non con cellule normali. Era come se l'HTLV-I fosse la causa di quella leucemia.

Una conferma di questa conclusione è stato offerta nel marzo del 1981 al corso pratico sui virus di Kyoto, dove ho presentato i miei primi risultati. Yorio Hinuma dell'Università di Kyoto ha descritto una linea cellulare colpita da ATL e prodotta da Isao Miyoshi dell'Università di Kochi. Questa linea, coltivata in laboratorio, liberava particelle retrovirali. Tutti i dati a disposizione indicavano che il virus proveniente dalle cellule di Miyoshi era identico all'HTLV-I.

Non molto tempo dopo questo punto è stato ribadito. Mitsuaki Yoshida dell'Istituto nazionale del cancro a Tokyo ha determinato la sequenza di basi nucleotidiche nel genoma a RNA del virus isolato dal suo compatriota e l'ha confrontata con la sequenza dell'HTLV-I. La sovrapposizione tra le due sequenze era sufficiente a far concludere che i due virus isolati, quello ottenuto dai giapponesi e quello ottenuto dagli americani, erano in realtà due ceppi di un unico virus, uniti da un'affinità molto stretta. La prima conferma indipendente della presenza di HTLV-I in pazienti statunitensi è stata data in seguito da Barton Haynes e Dani Bolognesi del Medical Center della Duke University, i quali hanno isolato il virus da un paziente statunitense di origine giapponese, affetto da una leucemia a linfociti *T*.

La concentrazione dell'ATL nel Giappone meridionale aveva contribuito a collegare l'HTLV-I a una specifica malattia; i miei collaboratori e io speravamo di identificare analoghi schemi anche altrove. Fino al 1980, però, la nostra ricerca epidemiologica fu gravemente rallentata dalla mancanza di aree di concentrazione della malattia, chiaramente definite, negli Stati Uniti. L'unico dato epidemiologico saliente era che, per la maggior parte, i pazienti che avevamo in cura erano negri nati negli Stati Uniti, nei paesi caraibici o nell'America Meridionale. Nel 1981, Daniel Catovsky dell'Hammersmith Hospital di Londra segnalò una concentrazione di leucemie a linfociti *T* nei negri nati nei Caraibi. Gli aspetti clinici rilevati erano notevolmente simili a quelli dei casi che erano stati curati negli Stati Uniti e in Giappone e che avevano mostrato una positività per l'HTLV-I; come ormai ci aspettavamo, l'HTLV-I fu trovato in tutti.

L'intuito di Catovsky fornì una delle chiavi che hanno permesso di sbloccare l'epidemiologia dell'HTLV-I a livello internazionale. Con la validissima collaborazione di William A. Blattner del National Cancer Institute e dei suoi colleghi, i miei collaboratori Carl Saxinger e Marjorie Robert-Guroff e io abbiamo cominciato ad analizzare le popolazioni negre degli Stati Uniti, dei Caraibi, dell'America Meridionale e dell'Africa per quanto riguarda le infezioni da HTLV-I. Abbiamo così trovato che questo virus è endemico non soltanto nelle isole meridionali del Giappone, ma anche in alcune zone degli Stati Uniti, nella maggior parte dei Caraibi, nelle regioni settentrionali dell'America meridionale e, in particolare, in Africa.

Commercio e cancro

Che cosa univa queste regioni così disparate? La risposta, sufficientemente esauriente, è il mercato degli schiavi,

un'ipotesi che però non emerse direttamente. Miyoshi aveva rilevato che alcuni macachi giapponesi possedevano anticorpi contro l'HTLV-I e suppose che in Giappone il virus avesse infettato l'uomo a partire da tali scimmie. È stato però dimostrato che ciò era falso: il virus dei macachi giapponesi è diverso dall'HTLV-I in maniera tale da permettere di porre fuori causa la trasmissione diretta. Nel confermare i risultati ottenuti da Miyoshi, Gebhard Hunsman dell'Università di Gottinga e il mio gruppo hanno, tuttavia, scoperto che molte specie di scimmie africane producono anticorpi che reagiscono con l'HTLV-I. I virus poi isolati da quelle scimmie erano anch'essi affini all'HTLV-I, benché distinti.

Alcuni virus africani, in particolare quelli dei cercopitechi grigioverdi e degli scimpanzé, sono risultati molto più affini all'HTLV-I del virus del macaco giapponese. La rassomiglianza era sufficiente a suggerire una stretta connessione tra l'infezione di quelle specie e l'infezione degli esseri umani. Sulla base di questi risultati (assieme a dati epidemiologici e a informazioni di natura storica) ho proposto la seguente ipotesi. L'HTLV-I ebbe origine in Africa, dove infettò molte specie di primati del Vecchio Mondo, tra cui gli esseri umani, e poi raggiunse le Americhe con il mercato degli schiavi.

Fatto curioso, l'HTLV-I potrebbe essere arrivato in Giappone per la stessa via. Nel XVI secolo, i mercanti portoghesi sbarcarono in Giappone e soggiornarono proprio in quelle isole dove oggi l'HTLV-I è endemico. Portarono con sé anche scimmie e schiavi africani, come testimoniano opere d'arte giapponesi dell'epoca, e o le une o gli altri potrebbero essere stati i vettori del virus. Questa ipotesi è stata messa in discussione recentemente quando è stato scoperto che l'infezione da HTLV-I è comune tra gli Ainu che vivono nell'isola di Hokkaido, la più settentrionale delle grandi isole del Giappone, mai sfiorata dai portoghesi. Tuttavia, per il momento essa rimane una spiegazione plausibile del quadro globale della diffusione.

I dati sull'epidemiologia dell'HTLV-I hanno chiarito non solo il quadro complessivo, ma anche i diversi quadri in aree più ristrette e perfino i meccanismi mediante i quali il virus viene trasmesso da una persona a un'altra. Il legame fondamentale tra HTLV-I e l'ATL era relativamente semplice, in quanto il virus presenta una diffusione maggiormente localizzata rispetto ad altri virus e la corrispondenza tra virus e malattia è marcata: dove si trovano concentrazioni di ATL, il virus è predominante e viceversa. In piccole aree geografiche come i distretti, tuttavia, l'infezione prevale in modo molto variabile.

Questa notevole variazione all'interno di aree ristrette (assieme ad altri dati) mostra che l'HTLV-I non viene trasmesso casualmente. Sono state stabilite parecchie vie di infezione e tutte necessitano di scambi stretti. Il virus può essere trasmesso nelle trasfusioni di sangue contaminato o tra drogati che usino la stessa siringa. Può anche essere trasmesso mediante contatto sessuale, sia etero- sia omosessuale. Un feto può essere infettato nell'utero se la madre ospita il virus. Recenti osservazioni effettuate in Giappone suggeriscono che i lattanti possono infettarsi con l'HTLV-I succhiando il latte materno. Dati ancora più recenti indicano che il virus può essere veicolato dalle zanzare. Una volta avvenuta la trasmissione, subentra un periodo di latenza che può durare anche 40 anni (se l'infezione ha luogo nell'infanzia) oppure solo pochi anni (se avviene nell'adulto).

I mercanti portoghesi potrebbero aver portato l'HTLV-I in Giappone dall'Africa mediante il commercio di schiavi o di scimmie. Nel XVI secolo i navigatori portoghesi si recarono numerose volte in Giappone e l'illustrazione mostra un dipinto dell'epoca che rappresenta una di quelle visite. A sinistra si nota un negro che regge un baldacchino sopra due mercanti portoghesi impegnati in un'animata conversazione. I contatti con i portoghesi si concentrarono nelle isole più meridionali del Giappone dove, nel 1978, è stato identificato un nuovo tipo di neoplasia a linfociti *T*. Per la malattia nota con il nome di leucemia a linfociti *T* dell'adulto o ATL è stato dimostrato il collegamento con l'HTLV-I.

La causa diretta

Qual è la relazione tra l'infezione originaria e il cancro che può comparire anche 40 anni dopo? Diversi filoni di testimonianze suggeriscono che il ruolo del virus è diretto. Un filone consiste di dati epidemiologici. Bambini infetti, nati nell'area endemica del Giappone meridionale, hanno la stessa probabilità di sviluppare l'ATL se trascorrono il resto della loro esistenza in quella stessa zona oppure se si trasferiscono, precocemente negli anni, in altra parte del mondo. Lo stesso vale per altre aree endemiche. Sembra che il virus possa avviare la catena di eventi che conduce al cancro da solo e che non siano necessari fattori ambientali presenti nell'area endemica.

Il secondo filone di testimonianze proviene dalla virologia molecolare. Quando l'HTLV-I infetta i linfociti *T* in laboratorio, il suo provirus si integra a caso nel genoma di queste cellule: in un linfocito può trovarsi sul cromosoma 5, in un altro sul cromosoma 8 e in un terzo ancora sul cromosoma 15. Eppure, quando vengono esaminati i linfociti di un paziente affetto da ATL, le sequenze virali si trovano allo stesso posto in ogni linfocito leucemico. Questa identità implica che la neoplasia sia in realtà un clone: tutte le sue cellule sono progenie di una singola cellula capostipite e come tali contengono copie del genoma della progenitrice. Implica anche che l'infezione ha preceduto l'origine della neoplasia, perché se il virus fosse penetrato nelle cellule di una neoplasia preesistente, le sequenze virali si sarebbero dovute trovare in una sede diversa in ogni cellula neoplastica.

E non è tutto. Quando Miyoshi ha coltivato cellule affette da ATL con sangue di cordone ombelicale umano, i linfociti *T* del sangue del bambino (normale) si sono trasformati entrando in una condizione precancerosa, a vita illimitata. Probabilmente le particelle virali, sfuggite dalle cellule leucemiche, avevano trasformato quei linfociti normali. Questa spiegazione è stata confermata in seguito, in un modo inequivocabile, dagli esperimenti iniziati nel mio laboratorio da Mika Popovic e continuati da Zaki Salahuddin e Philip Markham, i quali hanno mostrato che molti HTLV-I isolati possono trasformare i linfociti *T4* in un arco di tempo che va dalle cinque alle sette settimane. Il provirus delle cellule trasformate è integrato secondo uno schema analogo a quello notato nelle cellule colpite da ATL. Il cerchio è stato chiuso da Miyoshi, che ha preso cellule trasformate in laboratorio, le ha introdotte in criceti e ha mostrato che sono in grado di sviluppare cancro.

L'avere collegato l'HTLV-I direttamente a una forma tumorale maligna è stato stimolante, ma ben difficilmente avrebbe potuto rappresentare un punto di arrivo. Al contrario, ha costituito il punto di partenza per il traguardo successivo, che è consistito nell'identificare il meccanismo molecolare mediante il quale il virus provoca la leucemia. Anche se il programma non è stato completato, dalla ricerca finora svolta sono emerse alcune linee di sviluppo del processo patologico. Uno dei chiarimenti fondamentali è venuto dall'esame dello schema di integrazione del provirus nel genoma della cellula (uno schema elaborato dai miei collaboratori Flossie Wong-Staal, Beatrice Hahn e Vittorio Manzari e, indipendentemente da loro, da Motoharu Seiki dell'Istituto nazionale del cancro a Tokyo, che lavorava con Yoshida). Come ho già ricordato, in ogni cellula di un particolare tumore il provirus si integra alla maniera di un clone. Tuttavia, da un paziente affetto da ATL all'altro non c'è uniformità: ciascuno presenta le sequenze virali in una posizione diversa.

I retrovirus che provocano leucemia possono essere suddivisi in tre gruppi in base alla struttura genetica e al meccanismo patogenetico. La parte fondamentale del genoma retrovirale include sequenze di DNA che codificano per le proteine della parte centrale, per la transcriptasi inversa e per le proteine dell'involucro: esse sono note come *gag*, *pol* ed *env*. I virus che provocano la leucemia cronica, come il virus della leucemia murina (MuLV), hanno solo queste tre sequenze. I virus della leucemia acuta, o virus dei sarcomi, hanno in più un'altra sequenza: un gene cellulare, noto come *onc*, interessato nella regolazione della crescita cellulare. Tra i virus dei sarcomi si ha il virus del sarcoma aviario (ASV). Il terzo gruppo, dei virus che regolano per azione in *trans*, comprende l'HTLV-I e il suo affine HTLV-II ed è definito dalla presenza di un gene, il *tat*.

Un meccanismo nuovo

La combinazione tra capacità di formare cloni e casualità dell'integrazione da tumore a tumore ha subito messo in rilievo che l'HTLV-I deve provocare la leucemia mediante un meccanismo nuovo. Prima della scoperta dell'HTLV-I si conoscevano soltanto due meccanismi mediante i quali un retrovirus può provocare la leucemia e ciascuno di essi servì a definire un gruppo di virus. Un meccanismo comportava la presenza, nel genoma virale, di un gene *onc*, un gene cellulare coinvolto nella regolazione della crescita. L'espressione di questo gene provoca la trasformazione di ogni cellula infetta, il che dà origine a un cancro che si sviluppa rapidamente e che (provenendo da molti differenti progenitori) non è monoclonale. Questo meccanismo è comune a una varietà di virus della leucemia acuta e di virus responsabili di sarcomi (tra cui il virus del sarcoma aviario isolato da Rous). Tuttavia l'evento genetico che conduce alla combinazione necessaria tra DNA cellulare e DNA virale è raro e questi virus hanno scarsa rilevanza generale.

Molto più comuni sono i virus della leucemia cronica, tra cui il virus della leucemia aviaria (AVL) e il virus della leucemia murina (MuLV), che sono privi entrambi del gene *onc*; pertanto, la malattia che provocano è molto più simile alle neoplasie naturali umane: solo una minoranza degli animali infettati sviluppa leucemia; le leucemie compaiono solo dopo una lunga latenza; esse sono cloni. Ma tra il loro meccanismo d'azione e quello dell'HTLV-I vi è una differenza fondamentale: perché questi virus possano causare la malattia, il provirus deve integrarsi in un punto specifico, autorizzando così le sequenze regolatrici, presenti nel DNA virale, a interagire con i geni cellulari vicini, che promuovono la proliferazione delle cellule. D'altra parte, l'HTLV-I è chiaramente patogeno e non ha importanza dove, nel genoma, il suo provirus si integri.

Oltre a indicare l'indirizzo della ricerca futura, il contrasto tra i virus della leucemia cronica e l'HTLV-I ha aiutato a chiarire un vecchio rompicapo. Come ho già sottolineato, in buona parte il dubbio che mai retrovirus umani sarebbero stati trovati derivava dal fatto che la microscopia elettronica non riusciva a individuare particelle virali in cellule tumorali umane. Allorché il meccanismo d'azione dell'HTLV-I fu svelato, venne alla luce anche una possibile ragione per il basso tenore di virus nelle cellule neoplastiche. Questo meccanismo richiede la presenza di poche particelle virali perché l'HTLV-I può innescare la catena di eventi che conduce alla comparsa di un tumore integrandosi in un punto qualsiasi del ge-

La proteina codificata dal gene *tat* svolge funzioni di cruciale importanza nella replicazione dell'HTLV-I e nella trasformazione della cellula ospite in uno stato canceroso. Il gene *tat* è situato vicino a un'estremità del provirus, subito dopo un segmento di DNA chiamato unità ripetitiva terminale o LTR (*1*). Le due LTR regolano l'attività di altri geni virali. Legandosi a esse, la proteina codificata dal gene *tat* aiuta a controllare la velocità della replicazione virale (*2*); unendosi alle sequenze regolatrici cellulari (che controllano i geni per la IL-2 e per il suo recettore), induce invece la cellula ospite a proliferare anormalmente, il che è forse il primo passo verso il cancro.

noma cellulare. Al contrario, il meccanismo d'azione dei virus della leucemia cronica dipende dall'integrazione del virus stesso in una posizione specifica e l'abbondante replicazione potrebbe essere un modo per assicurare che almeno una copia del provirus raggiunga la posizione esatta.

Azione a distanza

È chiaro che vi sono differenze significative tra i modi d'azione dei due tipi di virus. Eppure si ammette, in generale, che entrambi i meccanismi finiscano con l'influire sull'attività dei geni cellulari che controllano la proliferazione. Il provirus dei virus che causano la leucemia cronica lo fa perché è vicino a quelle sequenze. D'altra parte, il provirus dell'HTLV-I può essere integrato in un punto qualsiasi del genoma della cellula e, se deve attivare i geni che promuovono la crescita, deve farlo agendo a distanza. In biologia molecolare si dice che questo tipo di meccanismo agisce in *trans*. Qual è questo meccanismo di azione in *trans* nel caso dell'HTLV-I?

Il primo barlume di risposta si è avuto quando Seiki e Yoshida hanno determinato la sequenza nucleotidica del provirus e hanno trovato una regione, nuova tra tutti i retrovirus noti, che hanno chiamato *X*. Essi hanno notato che *X* poteva codificare in teoria per quattro proteine. Flossie Wong-Staal e il mio collaboratore George Shaw, assieme a William Haseltine del Dana-Farber Cancer Institute, hanno identificato poco dopo un segmento di quell'*X*, che è risultato essere il gene che codifica per una delle quattro proteine. Il gene, comune a tutti gli HTLV-I isolati, è stato chiamato *tat*, il che significa un gene responsabile dell'attivazione per azione in *trans*. Da allora è stato chiaro che la proteina codificata dal gene *tat* è essenziale perché il virus possa esplicare in coltura i suoi effetti trasformanti.

Un indizio su che cosa il gene *tat* poteva fare è emerso da alcuni primi tentativi di coltivare i linfociti *T*. Come ho già descritto, avevamo trovato che alcuni linfociti *T* neoplastici potevano essere coltivati con IL-2 senza alcuna precedente attivazione. Di solito l'IL-2 è scarsa nell'organismo e i geni per i suoi recettori sono inattivi. I nostri risultati hanno, tuttavia, suggerito che i geni per l'IL-2 potevano essere attivati e mantenuti in questo stato nei linfociti neoplastici. In un recente affascinante esperimento, il mio collega Warner Greene ha scoperto che, quando il gene *tat* è inserito nei linfociti *T* umani, questi non solo producono IL-2, ma cominciano anche a produrre i suoi recettori. Sembra probabile che, in una fase precoce della trasformazione, il gene *tat* induca la produzione sia di IL-2 sia dei suoi recettori; i linfociti che hanno subito questa duplice induzione vanno quindi incontro a una crescita anomala.

Eppure devono intervenire altri eventi lungo il percorso che conduce alla condizione neoplastica perché, dal momento che un paziente ha completamente sviluppato l'ATL, le sue cellule leucemiche né producono più IL-2, né hanno bisogno di essa. Inoltre, allo stadio in cui la leucemia è marcata, nessun gene virale, neppure il *tat*, si esprime nella cellula neoplastica a livelli riconoscibili. È probabile che dopo l'avvio della trasformazione da parte del *tat*, la cellula infetta proliferi abbondantemente. Ma è necessario un secondo (e forse un terzo) evento genetico per produrre in alcune cellule discendenti lo stato neoplastico finale.

Fatto interessante, il prodotto del gene *tat* svolge un ruolo nella replicazione virale, come pure nel promuovere la degenerazione neoplastica delle cellule. A ognuna delle due estremità del provirus esiste un segmento di DNA, formato dalle cosiddette unità ripetitive terminali o LTR (*Long Terminal Redundancy*). Le LTR includono sequenze che regolano l'espressione dei geni virali e pertanto determinano la velocità alla quale sono prodotti i componenti del virus (RNA e proteine). Interagendo con alcune di queste sequenze regolatrici, la proteina codificata dal gene *tat* può fare aumentare rapidamente il livello di produzione di nuove particelle virali. Questo fatto può rappresentare un notevole beneficio per l'HTLV-I, che si replica soltanto durante una limitata porzione del ciclo vitale della cellula (cioè nel periodo in cui vengono attivati i linfociti *T*) e probabilmente ha bisogno di sfruttare al massimo questa breve parentesi.

Altre vie

Questa è, per il momento, una sintesi in breve delle nozioni che si hanno sui meccanismi d'azione dell'HTLV-I a livello molecolare. Mentre la ricerca procede in questa direzione vengono seguite anche numerose altre vie di indagine. Una riguarda la gamma di malattie con le quali il virus è associato. Oltre a essere la causa diretta di neoplasie a linfociti *T* negli adulti, l'HTLV-I può essere un fattore che contribuisce indirettamente a determinare altre condizioni patologiche.

Per esempio, Essex ha suggerito che pazienti infettati con il virus vadano soggetti ad altre infezioni, forse perché alcuni linfociti *T* infetti, anche se non trasformati, hanno subito un danno funzionale. Anche alcuni linfociti *T* «normali» infetti possono contribuire a determinare leucemie a carico di un altro tipo di globuli bianchi, i linfociti *B*, che si osservano in alcuni pazienti di zone endemiche dei Caraibi. (La connessione tra linfociti *T* infetti e leucemie a linfociti *B* può indicare la via verso una più approfondita comprensione del modo in cui la leucemia si sviluppa.) Infine, recenti ricerche hanno mostrato che l'HTLV-I è associato a una malattia neurologica che ricorda la sclerosi multipla cronica. Sembra evidente che l'impatto globale dell'HTLV-I sulla salute pubblica stia solo ora cominciando a essere riconosciuto.

Le ricerche stanno procedendo anche sul secondo retrovirus umano, l'HTLV-II, che è stato scoperto nel 1982 dal mio gruppo in uno studio effettuato in collaborazione con David W. Golde della School of Medicine dell'Università della California a Los Angeles. Il primo HTLV-II isolato proveniva da un ceppo di linfociti *T* prelevato da un giovane bianco, affetto da una variante di una malattia a carico dei linfociti *T*, nota come tricoleucemia. L'HTLV-II differisce in maniera sottile per forma e funzione dal virus cugino ed è associato a forme di leucemia a linfociti *T* meno aggressive. Tuttavia, negli aspetti più significativi i due virus sono simili (come è stato dimostrato da Golde e dal suo collaboratore Irvin Chen oltre che dai miei collaboratori Wong-Staal ed Edward Gelman). Complessivamente l'HTLV-I e l'HTLV-II hanno la stessa struttura genomica e nella loro sequenza nucleotidica vi è una considerevole sovrapposizione. I due virus mostrano la stessa capacità di trasformare le cellule in coltura. Inoltre hanno in comune lo stesso meccanismo d'azione in *trans* e un prodotto del gene *tat*, designato come *tat* II, è stato trovato nel virus scoperto più di recente.

Le malattie provocate dall'HTLV-I e dall'HTLV-II sono piuttosto rare (tranne che nelle zone endemiche). Potrà non essere sempre così. Vi sono prove che entrambi i virus si stiano diffondendo in alcune popolazioni che si trovano a un livello maggiore di rischio d'infezione e, in queste popolazioni, particolarmente tra coloro che fanno uso di droga per via endovenosa. Preoccupata da questi risultati, la Croce Rossa sta conducendo un'indagine per determinare se occorra sottoporre a screening, riguardo ai due virus, il sangue dei donatori, così come è già stato fatto per l'HTLV-III. Al momento in cui questo articolo sarà pubblicato, probabilmente una decisione sarà stata presa e ci sono molte probabilità che essa sia positiva.

La nostra sempre più ricca conoscenza dei primi retrovirus umani è il risultato di una ricerca cominciata fin dal 1910. In questa lunga storia, un episodio - l'isolamento dell'HTLV-I e il suo legame con una forma neoplastica nell'uomo - ha avuto importanza critica. Quest'impresa si basava su due condizioni preliminari: la prima, un saggio sensibile per dimostrare la presenza di virus, è stata resa possibile dalla scoperta della transcriptasi inversa; la seconda si è potuta realizzare grazie alla definizione di un metodo per coltivare in laboratorio i linfociti *T*. Senza un metodo per coltivarli, non ci sarebbe potuta essere alcuna prova sul sangue, né anticorpi monoclonali per le indagini epidemiologiche o sonde di DNA per capire i meccanismi molecolari del virus. Queste tecniche formano la base di un metodo che ha assunto massimo significato quando il mondo fu colpito dalla prima grande pandemia della seconda metà del XX secolo: l'AIDS.

La materia oscura dell'universo

Strettamente legata all'origine dell'universo, alla cui massa contribuisce per circa l'80 per cento, è al centro di numerose ricerche e speculazioni sia nel campo dell'astrofisica sia in quello della fisica delle alte energie

di Lawrence M. Krauss

Di che cosa è fatto l'universo? Qual è il tipo di materia più diffuso, qual è la sua entità e come si distribuisce? Queste domande, che sono sempre state al centro dell'interesse dei cosmologi, da qualche anno si sono fatte ancora più pressanti, a mano a mano che si accumulano le indicazioni a favore dell'idea secondo la quale la maggior parte della massa dell'universo sarebbe oscura, invisibile a qualunque telescopio o ad altri mezzi di osservazione, e che gli ultimi sviluppi sia della fisica delle alte energie sia dell'astrofisica rendono possibili nuove previsioni sulla composizione e la distribuzione di questa forma di materia probabilmente diversa dalla materia comune.

Vi sono già moltissime prove che la materia visibile contenuta nelle galassie può contribuire per meno del 10 per cento alla massa effettiva delle galassie mentre il resto, non ancora individuabile direttamente, è distribuito probabilmente all'interno e intorno a ciascuna galassia. Oggi però alcune considerazioni teoriche inducono a ritenere che questa sia forse solo la punta dell'«iceberg» cosmico della materia oscura: può darsi che in tutto l'universo ne siano distribuite quantità molto maggiori, magari in configurazioni completamente indipendenti dalla distribuzione delle galassie. È possibile che questa massa sia suscettibile di spiegazione solo invocando l'esistenza di nuovi tipi di materia.

Il problema della materia oscura, il problema cioè della sua quantità, della sua distribuzione e della sua composizione, è profondamente legato a quelli che riguardano la struttura e l'evoluzione complessiva dell'universo: dal momento che è probabilmente la forma di massa che predomina nell'universo, la materia oscura deve avere influenzato l'evoluzione delle caratteristiche oggi osservabili. I problemi relativi alla struttura rinviano a loro volta, per essere spiegati, allo stretto legame che si è venuto a formare tra macrofisica e microfisica, i due corpi di conoscenza che descrivono rispettivamente le interazioni di scala più grande (ossia quella relativa all'universo nel suo complesso) e quelle di scala più piccola (ossia quella delle particelle fondamentali che costituiscono tutta la materia).

Il legame è dato dall'osservazione che l'universo si sta espandendo. Se si ha il coraggio di estrapolare nel passato questa espansione per 10 o 20 miliardi di anni, le due scale cominciano a fondersi, perché nei primissimi istanti di vita dell'universo quelle strutture che oggi si osservano alle scale più grandi occupavano regioni con distanze ed energie caratteristiche di scala tipica dei processi che governano le interazioni tra le particelle fondamentali. Poiché la struttura residua alle scale più grandi oggi osservabili riflette l'impronta di quei processi, è logico attendersi che la soluzione del problema della materia oscura venga in parte dai progressi nella conoscenza della fisica delle particelle di alta energia.

A tutt'oggi sono state prospettate numerose previsioni verificabili sulla natura tanto della materia oscura quanto delle strutture primordiali esistenti nelle prime fasi della vita dell'universo. Gli sviluppi futuri, teorici e osservativi, contribuiranno a dirimere numerose questioni, da quelle sulla modalità e sull'epoca di formazione delle prime galassie e delle prime stelle a quella sui tipi di simmetria alla base delle interazioni tra particelle di altissima energia. Alla fine può darsi che il dibattito sulla materia oscura aiuti a trovare la risposta a una domanda che è vecchia quanto la curiosità umana: quale sarà il destino dell'universo?

Fin dall'inizio degli anni trenta, quando Edwin P. Hubble confermò che l'universo si sta espandendo, è risultato naturale chiedersi se questa espansione alla fine si arresterà. La risposta dipende da due fattori: l'attuale velocità di espansione dell'universo e l'intensità, determinata dalla densità media di massa entro l'universo, della forza gravitazionale che tiene insieme questa massa. Una densità di massa elevata causerebbe una forte attrazione gravitazionale.

Secondo la teoria della relatività generale esiste un rapporto tra il valore della velocità di espansione e quello della densità di massa da una parte e la curvatura media dell'universo dall'altra (*si veda l'illustrazione a pagina 36*). Se la densità media di massa è abbastanza elevata da arrestare l'espansione e far contrarre di nuovo l'universo, questo si dice chiuso. Se invece la densità è tanto piccola rispetto alla velocità di espansione che l'universo continuerà a espandersi con velocità finita per sempre, l'universo si dice aperto. Qualora infine l'attrazione gravitazionale possegga l'intensità esattamente necessaria per continuare a rallentare l'espansione ma insufficiente a chiudere l'universo, questo si dice piatto.

Dato che l'universo osservabile è estremamente uniforme in tutte le direzioni, la sua velocità di espansione si può descrivere in termini di un solo parametro, la costante di Hubble, costante per tutto lo spazio in ogni momento dato, ma lentamente variabile nel tempo. La costante di Hubble è pari alla velocità media alla quale due regioni qualsiasi dell'universo si allontanano l'una dall'altra divisa per la distanza che le separa. Per un qualunque valore misurato della costante di Hubble è facile stabilire la densità di massa che corrisponderebbe a un universo piatto. Le misurazioni di questo parametro, però, dipendono da quelle di una serie di altri poco noti. Di solito la costante di Hubble si determina misurando la velocità alla quale si allontanano dalla Terra vari oggetti, e valutandone con la massima precisione possibile la distanza mediante tecniche quali il confronto tra la loro brillanza intrinseca, stimata, e quella apparente vista dalla Terra.

Dato però che queste misurazioni sono

La galassia a spirale M31 (la nebulosa di Andromeda) rivela la presenza di materia oscura grazie al moto dei bracci esterni, che ruotano intorno al centro galattico più rapidamente di quanto ci si aspetterebbe se la materia visibile, luminosa, della galassia ne costituisse tutta la massa o quasi. Una quantità di materia circa 10 volte maggiore si trova probabilmente distribuita in una grande sfera di materia oscura, che avvolge completamente la galassia luminosa. I risultati delle osservazioni, uniti a considerazioni di ordine cosmologico, inducono a ritenere che la massa associata alle galassie, compresa quella della materia oscura nella quale sono immerse, contribuisca solo per il 20 per cento circa della densità di massa totale dell'universo. Il resto potrebbe essere associato a materia oscura di natura inconsueta distribuita altrove.

assai poco precise, le valutazioni attuali della velocità di espansione dell'universo variano di circa un fattore due. Considerando il limite superiore, oggetti distanti un megaparsec (circa 3,26 milioni di anni luce) si stanno allontanando in media a una velocità leggermente inferiore a circa 100 chilometri al secondo. A questa velocità la densità media di massa che darebbe luogo a un universo piatto è di circa 2×10^{-29} grammi per centimetro cubo, circa la massa di 10 atomi di idrogeno per metro cubo di spazio.

Come è possibile stabilire quanta massa esiste in realtà? Un metodo per trovare almeno un limite inferiore è semplicemente quello di sommare la quantità totale di materia visibile. Dal momento però che quella suscettibile di misurazione diretta non è la massa ma la luminosità, occorre un certo lavoro di

La curvatura dell'universo dipende dalla sua velocità di espansione media e dalla densità media della materia che contiene. Se la velocità di espansione è elevata rispetto alla quantità di massa l'universo è aperto. La sua geometria spaziale in tre dimensioni è analoga a una geometria bidimensionale particolare, l'area presso il centro di una sella (*a, a sinistra*). La forma dello spazio si ripercuote su quella delle figure geometriche. La somma degli angoli di un triangolo, per esempio, in un universo aperto sarebbe inferiore a 180 gradi, proprio come accade sulla superficie di una sella. L'effetto risulterebbe impercettibile su distanze piccole come quelle misurabili sulla Terra, proprio come un triangolo su una regione molto piccola della sella sembrerebbe normale. Se l'universo è aperto continuerà a espandersi con velocità finita e non nulla per sempre: qualsiasi lunghezza di riferimento (la distanza tra due regioni qualunque dell'universo in espansione) continuerà a crescere (*a, a destra*). Se invece la quantità di massa nell'universo è grande rispetto alla velocità di espansione, l'universo è chiuso, cioè si chiude su se stesso, in modo analogo a quello della superficie di una sfera (*b, a sinistra*). Su una sfera, come in un universo chiuso, la somma degli angoli di un triangolo supera i 180 gradi. Se l'universo ha questa forma alla fine smetterà di espandersi e si contrarrà nuovamente (*b, a destra*). Se infine la velocità di recessione e la quantità di massa presente nell'universo si compensano esattamente, l'universo risulta piatto, analogo dal punto di vista geometrico a un piano (*c, a sinistra*). Continuerà a espandersi, ma la velocità di espansione calerà asintoticamente (*c, a destra*). Esistono svariate considerazioni teoriche di notevole validità a favore dell'ipotesi secondo la quale l'universo in realtà sarebbe piatto, anche se, perché sia così, è necessario che contenga molta più materia di quanta non ne sia stata osservata finora, direttamente e indirettamente.

interpretazione per tradurre i risultati delle osservazioni in densità di massa presunte. Combinando la distribuzione e la luminosità osservate degli oggetti stellari e del gas diffuso con le valutazioni teoriche della loro massa sembra che il valore del rapporto tra massa e luminosità della materia fotoemittente legata alle galassie sia alcune volte superiore a quello del Sole. Partendo da questa valutazione e dai limiti inferiori stimati per la costante di Hubble si ricava che la densità media della materia luminosa nell'universo è inferiore al due per cento circa di quella necessaria per arrestare l'espansione dell'universo.

Sin dal 1933, però, si sa come gli ammassi di galassie contengano una percentuale notevole di massa non luminosa. In quell'anno infatti Fritz Zwicky del California Institute of Technology, che stava analizzando le velocità delle singole galassie appartenenti all'ammasso in Coma, scoprì come molte di esse si muovano così velocemente che l'ammasso dovrebbe tendere a disperdersi se non vi fosse a tenerlo insieme più massa di quella luminosa solamente. Altre considerazioni però indicavano che l'ammasso era stabile, e così Zwicky concluse che doveva contenere materia non luminosa.

Dimostrando come in linea di principio fosse possibile rilevare indirettamente la materia oscura tramite i suoi effetti gravitazionali, Zwicky creò un precedente di grande importanza. Negli ultimi anni i ricercatori hanno dimostrato in maniera convincente che tecniche simili possono rilevare la presenza di materia oscura in strutture di scale che vanno da quella dei dintorni immediati del Sole a quelle delle galassie e degli ammassi di galassie, fino a quella dei superammassi costituiti da migliaia di galassie.

L'indicazione meglio documentata della presenza di materia oscura si basa sulla velocità di rotazione delle galassie a spirale (si veda l'articolo *La materia oscura nelle galassie a spirale* di Vera C. Rubin in «Le Scienze» n. 180, agosto 1983). La velocità di rotazione di un oggetto in un sistema stabile legato dalla gravitazione, come una galassia a spirale, dipende tra l'altro dalla sua distanza dal centro di rotazione. Secondo le leggi di Newton, infatti, la velocità orbitale di oggetti che ruotano a grande distanza da una concentrazione di massa dovrebbe diminuire in proporzione inversa alla radice quadrata della distanza dal centro di rotazione. Nel corso di alcuni estesi rilevamenti di stelle e di gas caldo nelle regioni esterne delle galassie a spirale, diversi gruppi di ricerca hanno dimostrato che la velocità di rotazione di questi oggetti non diminuisce, ma rimane costante fino a distanze superiori a 30 chiloparsec dal nucleo galattico. Jeremiah P. Ostriker e P. James E. Peebles della Princeton University avevano già ipotizzato che dovesse esistere una certa quantità di massa sfuggita all'osserva-

Lo spostamento Doppler della frequenza permette di determinare la velocità con la quale un oggetto che emette luce si allontana o si avvicina rispetto a noi e quindi, in particolare, la velocità di rotazione dei bracci di una galassia a spirale. Un oggetto stellare emette luce a certe frequenze caratteristiche determinate dalla composizione chimica (*a*). Se l'oggetto si sta allontanando dall'osservatore (*b*), la lunghezza d'onda della luce osservata risulta aumentata. Il fenomeno prende il nome di spostamento verso il rosso perché la luce di lunghezza d'onda maggiore è più rossa. Se l'oggetto si muove verso l'osservatore (*c*), la lunghezza d'onda della luce diminuisce, cioè si ha uno spostamento verso il blu. Le galassie a spirale mostrano in prevalenza un moto di allontanamento dalla Terra a causa dell'espansione dell'universo. Per un osservatore sul nostro pianeta, quindi, la luce proveniente da una galassia a spirale risulta spostata verso il rosso (*d, al centro*). Uno dei due bracci della galassia in rotazione (*d, a sinistra*), però, si allontanerà dalla Terra a velocità inferiore a quella del centro galattico, e quindi la luce che emette presenterà uno spostamento verso il rosso inferiore. L'altro invece (*d, a destra*) si allontanerà dalla Terra più velocemente del centro della galassia e quindi la sua luce risulterà ancora più spostata verso il rosso. Confrontando lo spostamento verso il rosso del centro galattico e quello dei bracci si può stabilire la velocità di rotazione dei vari punti dei bracci e quindi la distribuzione di massa nella galassia.

zione nelle galassie a spirale, perché altrimenti le instabilità gravitazionali farebbero collassare le galassie in strutture allungate a forma di barra. La stabilità delle galassie a spirale, così come la velocità di rotazione dei loro bracci esterni, si potrebbe spiegare se si ammettesse che ognuna di esse sia immersa in una grande distribuzione approssimativamente sferica di materia oscura.

Esistono altre indicazioni di ordine dinamico a favore dell'esistenza di una massa oscura, sia su scala più grande, sia su scala più piccola di quella delle galassie singole. Queste indicazioni si ricavano non dalle misurazioni della velocità di rotazione, ma da quelle della velocità casuale dei singoli oggetti in moto all'interno di sistemi legati dalla gravitazione. Un teorema della meccanica classica, il teorema del viriale, sancisce un legame tra l'energia cinetica media e quella potenziale media di oggetti presenti in un sistema stabile legato dalla gravitazione che abbia raggiunto l'equilibrio dinamico. Dovrebbe essere quindi possibile valutare la massa totale di un sistema di questo genere (che è proporzionale alla sua energia gravitazionale totale) misurando la velocità relativa dei membri di un gran numero di coppie di oggetti appartenenti al sistema. Questo metodo ha fornito indicazioni sull'esistenza di materia oscura in una vasta gamma di sistemi, dalle galassie sferoidali nane di appena 10^7 masse solari fino agli ammassi di galassie che raggiungono 10^{15} masse solari. Sulle scale più grandi studiate con questo tipo di analisi la densità media di massa non supera il 20 per cento circa di quella necessaria per chiudere l'universo.

Un altro metodo, di cui Peebles e collaboratori sono stati pionieri, si basa sull'analisi statistica di grandi quantità di galassie invece che sui dati relativi a galassie o ammassi singoli. Peebles ha dimostrato che, accumulando dati statistici sul moto e l'aggregazione delle galassie a scale diverse e supponendo che la regione in esame contenga sistemi dinamici stabili per quanto riguarda la gravitazione, è possibile risalire dalla velocità relativa media di un gran numero di coppie di galassie alla densità media di massa dell'universo.

È notevole che tutti i metodi disponibili, non solo quelli discussi qui ma anche molti altri, diano sostanzialmente lo stesso risultato: se la distribuzione delle galassie è indicativa di quella della massa, allora l'universo non possiede neppure il 20-30 per cento circa della densità media di massa necessaria a chiuderlo.

Ma anche se le galassie non sono un indicatore valido della distribuzione della massa, o se per qualche motivo tutte le analisi risentono di errori sistematici, vi sono ancora valide ragioni per credere che in ogni caso la quantità di massa comune (cioè costituita prevalentemente da protoni e neutroni) dell'universo corrisponda a non più del 20 per cento circa di quella necessaria perché questo sia chiuso. Le indicazioni in questo senso provengono in prevalenza dal quadro teorico che spiega il processo di nucleosintesi nel corso del quale si sono formati per la prima volta vari elementi e isotopi leggeri che risultano abbondanti su scala cosmica.

La nucleosintesi degli elementi leggeri ha avuto luogo principalmente nei primi minuti dell'esistenza dell'universo. Il processo dovrebbe essere stato molto sensibile al valore della densità dei protoni e dei neutroni a quell'epoca. Se si vuole che le previsioni dei modelli teorici attuali della nucleosintesi concordino

con le abbondanze odierne degli elementi leggeri la densità totale dei protoni e dei neutroni che potevano essere presenti a quell'epoca è vincolata così strettamente che la loro densità attuale deve essere inferiore al 20 per cento circa di quella necessaria per la chiusura dell'universo. Sembra così che se l'universo è chiuso, almeno l'80 per cento della sua massa totale debba essere costituito da qualche altro tipo di materia.

Dato che simili considerazioni teoriche fondamentali limitano la quantità di massa comune nell'universo al 20 per cento della densità critica e dato inoltre che i dati osservativi inducono a ritenere che la quantità di massa oscura legata alle galassie e agli ammassi di galassie sia circa quella, perché i cosmologi non dovrebbero ammettere che l'universo in realtà sia aperto? Non è affatto impossibile immaginare una forma nella quale possa esistere senza essere visibile abbastanza materia comune da spiegare la dinamica delle galassie e degli ammassi: ma allora perché è necessario postulare l'esistenza di altre forme di massa? Perché esiste un problema della materia oscura al di là dell'accezione banale?

Vi sono due ostacoli teorici alla scelta dell'ipotesi più semplice, quella secondo la quale tutta o quasi la massa dell'universo sarebbe costituita da materia comune e la densità media sarebbe solo il 20 per cento di quella critica. Il primo ostacolo è posto da una combinazione della teoria della formazione delle galassie e di osservazioni del fondo di microonde che permea il cosmo.

È opinione comune che le galassie si siano formate alla fine di un processo di condensazione nel quale le regioni dell'universo più dense della media si contrassero sotto l'azione della forza di gravità fino a separarsi dal movimento espansivo di fondo formando sistemi legati isolati. Per circa 100 000 anni dopo il big bang, però, la materia comune non ebbe modo di condensare in questa maniera: era ancora troppo calda perché le particelle cariche costituenti si fossero combinate in atomi elettricamente neutri e quindi risultava composta da particelle cariche indipendenti. A causa di questo stato di ionizzazione, i moti microscopici in seno alla materia venivano profondamente influenzati dai campi di radiazione elettromagnetica di fondo: materia e radiazione erano cioè accoppiate. Le regioni di materia comune che vennero a essere più dense di quelle circostanti e più piccole delle dimensioni dell'orizzonte degli eventi (la massima distanza che la luce avrebbe potuto percorrere a partire dal big bang, e quindi la massima distanza attraverso la quale due sistemi fisici avrebbero potuto intrattenere nessi causali) non riuscirono a condensare ulteriormente, perché la pressione della radiazione ostacolava la forza attrattiva della gravità.

Alla fine l'universo divenne abbastanza freddo da permettere alle particelle di carica opposta di combinarsi, rendendo neutra la materia comune, che quindi si disaccoppiò dalla radiazione. Il bagno di radiazione di fondo termica al quale la materia era rimasta legata fino allora si ritrovò così libero di raffreddarsi via via che l'universo si espandeva e oggi costituisce la nota radiazione di fondo cosmica a microonde che riempie l'universo. Le osservazioni hanno dimostrato che questa radiazione è isotropa, cioè uguale in tutte le direzioni, fino a un grado di precisione molto elevato.

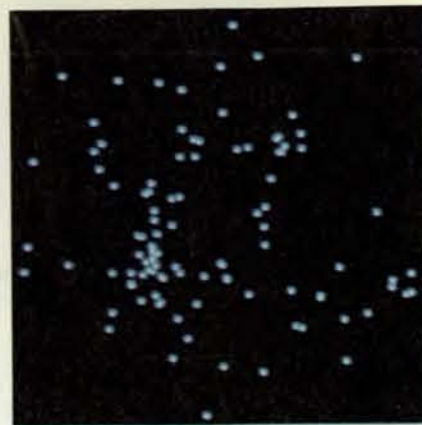
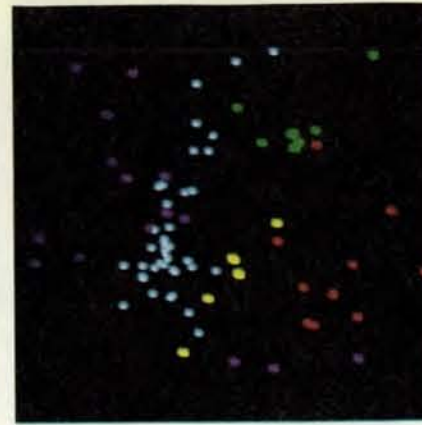
Poiché la gravità è una forza universalmente attrattiva, qualsiasi fluttuazione iniziale nella densità della materia comune dell'universo primordiale avrebbe presentato una tendenza a svilupparsi dopo che l'azione della pressione di radiazione smise di contrapporsi alla forza di gravità. Si suppone perciò che l'universo sia diventato e continui a diventare sempre più disomogeneo al passare del tempo e che le galassie, i cui nuclei hanno ora densità di oltre un milione di volte quella media del fondo, abbiano avuto inizio come fluttuazioni la cui densità era molto più vicina a quella del fondo.

Ma qual era l'entità delle fluttuazioni iniziali? A causa della limitatezza dei dati oggi disponibili sulle strutture di grande scala e delle difficoltà matematiche inerenti alla descrizione analitica dell'evoluzione di un sistema denso come una galassia, è estremamente difficile procedere a ritroso dallo stato attuale dell'universo fino a determinare la natura precisa delle prime fluttuazioni. Un metodo più facile è quello di supporre una struttura di fluttuazioni iniziali, di simularne lo sviluppo e l'evoluzione e di confrontare i risultati ottenuti con le osservazioni. Così facendo il cosmologo è guidato dai limiti sia superiori sia inferiori delle dimensioni delle fluttuazioni iniziali e dai vincoli sulla loro natura. In primo luogo devono essere state abbastanza intense perché le fluttuazioni di scala corrispondente alle dimensioni galattiche avessero tempo di condensarsi formando le galassie entro l'epoca attuale (in altre parole il rapporto fra eccesso di densità locale nella regione della fluttuazione e densità media nello spazio deve essere stato abbastanza grande da permettere che ciò avvenisse). In secondo luogo, le fluttuazioni dovevano presentare un'ampiezza abbastanza piccola da non lasciare nella radiazione di fondo un'anisotropia maggiore del limite superiore misurato.

Sembra però che queste due condizioni siano incompatibili l'una con l'altra se l'universo è composto prevalentemente da materia comune. Nel tempo compreso tra il disaccoppiamento della materia comune dalla radiazione e il momento in cui le fluttuazioni che avrebbero formato le galassie si contrassero formando sistemi isolati legati dall'attrazione gravitazionale, le fluttuazioni di densità, che inizialmente erano lievi, ebbero modo di crescere solo a una velocità ben precisa. Questa velocità era tale che quelle abbastanza grandi da avere avuto un tempo sufficiente per formare sistemi legati avrebbero portato a un'anisotropia della radiazione di fondo di oltre un ordine di grandezza maggiore dei limiti osservativi. In altri termini, tra il disaccoppiamento della materia comune e il momento attuale non vi è stato tempo sufficiente da permettere la formazione mediante la forza di gravità delle galassie a partire da variazioni di densità abbastanza piccole da non lasciare tracce osservabili nella radiazione di fondo.

Questa conclusione dipende da due ipotesi ampiamente sostenute, e precisamente quella secondo la quale il fondo a microonde non ha subito alterazioni degne di nota dopo il momento del disaccoppiamento e quella secondo cui è stata la sola gravità a portare alla formazione delle galassie. A meno che una o l'altra di queste due ipotesi classiche sia falsa, come sostengono vari ricercatori, sembra che sia necessaria qualche forma di materia nuova, capace di cominciare a condensare per gravità prima di quanto abbia potuto fare quella comune.

Esiste un secondo e fondamentale motivo per ritenere che nell'universo non predomini la materia comune con una densità pari a solo il 20 per cento circa della densità critica. Questo motivo, indicato come «problema della piattezza», è stato messo in evidenza per la prima volta da R.H. Dicke di Princeton e da Peebles. Secondo loro, il punto cruciale è che qualsiasi deviazione da un universo perfettamente piatto dovrebbe tendere a crescere linearmente con il tempo. Perciò se l'universo all'epoca della nucleosintesi avesse avuto una curvatura diversa da zero, anche piccola, entro oggi la deviazione sarebbe cresciuta di circa un fattore 10^{12}. Poiché però la densità della massa nell'universo attuale è vicina a quella di un universo chiuso entro un fattore 10 (in altri termini, poiché l'universo è relativamente prossimo a essere piatto), al momento della nucleosintesi l'universo deve essere stato o esattamente piatto o curvo in misura estremamente limitata, cioè piatto entro una parte su mille miliardi.

Se l'universo oggi presenta una curvatura misurabile, i cosmologi devono accettare il fatto miracoloso che ciò accade per la prima volta nei 10^{10} anni della sua storia, perché se una circostanza simile fosse avvenuta in un'epoca molto precedente oggi l'universo sarebbe curvo in modo molto più evidente di quanto non sia. Questa linea di ragionamento induce a ritenere che l'universo osservabile sia per sua stessa natura esattamente piatto, cioè che contenga esattamente la densità critica di massa. Poiché però la materia comune contribuisce probabilmente solo per il 20-30 per cento circa a questa densità, è probabile che esista qualche forma di materia di natura più inconsueta.

L'ammasso di galassie nel Cancro, come appare evidente da queste immagini al calcolatore, non è un sistema dinamico unico. Visto dalla Terra (*a sinistra*), sembra un complesso approssimativamente sferico in equilibrio apparente. Le misurazioni della velocità delle singole galassie al suo interno hanno però rivelato che il sistema tenderebbe a frantumarsi a meno che non contenga una grande quantità di materia non osservata. Un'analisi successiva di Gregory Bothun e colleghi dello Smithsonian Astrophysical Observatory ha però rivelato che in realtà l'ammasso è costituito da diversi gruppi di galassie distanziati (*al centro, nei vari colori*). Una vista ruotata, in uno spazio tridimensionale nel quale due assi rappresentano la posizione proiettata nel cielo e il terzo lo spostamento verso il rosso (*a destra*), mostra più chiaramente la separazione dei vari gruppi. Le velocità relative delle galassie all'interno di ciascun gruppo sono molto minori di quelle relative dei gruppi, una circostanza che indica come nel sistema esista complessivamente meno massa di quanto si fosse valutato in precedenza. Erano state proprio le velocità relative elevate dei gruppi a viziare gli studi precedenti. Questa analisi indica che la massa contenuta in ammassi di galassie come quello nel Cancro non è di per sé sufficiente a dar luogo a un universo piatto; deve di conseguenza esistere altra massa distribuita altrove. Queste immagini al calcolatore sono state realizzate da Michael J. Kurtz dello Smithsonian Astrophysical Observatory.

A questo punto è logico chiedersi perché l'universo sia esattamente piatto. Nel 1980 Alan H. Guth, ora al Massachusetts Institute of Technology, ha avanzato un'ipotesi di risposta in forma di un modello delle prime fasi dell'evoluzione dell'universo basato su alcune idee della fisica delle particelle formulate solo poco tempo prima.

Per la sua proposta Guth si è rifatto al lavoro di Howard Georgi e di Sheldon Lee Glashow della Harvard University, i quali nel 1974 avevano ipotizzato che tre delle forze fondamentali della natura, quella forte, quella debole e quella elettromagnetica, siano aspetti diversi di una sola forza «unificata». A energie sufficientemente elevate le tre forze dovrebbero essere esattamente simmetriche, cioè comportarsi in maniera identica, mentre a energie paragonabili a quelle oggi accessibili all'osservazione si possono comportare in modo molto diverso (si veda l'articolo *Una teoria unificata delle particelle elementari e delle forze* di Howard Georgi in «Le Scienze» n. 154, giugno 1981). La temperatura dell'universo subito dopo il big bang era abbastanza alta da permettere alla simmetria di queste tre forze di manifestarsi; quando l'universo si raffreddò al di sotto della temperatura critica alla quale era possibile conservare tale simmetria, la sua configurazione preferita divenne quella in cui la simmetria risultava «spezzata». L'effetto della rottura di questa simmetria fu che le forze assunsero un aspetto diverso l'una dall'altra.

Secondo l'idea di Guth, ampliata in seguito da Andrei D. Linde dell'Istituto di fisica P.N. Lebedev di Mosca e da Paul J. Steinhardt e Andreas Albrecht dell'Università della Pennsylvania, la rottura improvvisa della simmetria potrebbe aver fatto «gonfiare» rapidamente l'universo, in un'espansione esponenziale che lo avrebbe portato ad aumentare le proprie dimensioni di più di 28 ordini di grandezza in meno di 10^{-30} secondi. Dopo il periodo di espansione rapida, che dal vocabolo inglese corrispondente prende il nome di «inflazione», l'universo potrebbe essere tornato alla propria espansione normale, non esponenziale, che si osserva ancora oggi (si veda l'articolo *L'universo inflazionario* di Alan H. Guth e Paul J. Steinhardt in «Le Scienze» n. 191, luglio 1984).

Secondo questo modello, fu l'inflazione rapida dell'universo a far diventare piatte le regioni osservabili dell'universo, in maniera abbastanza simile a quanto avviene gonfiando un palloncino; dopo l'inflazione la parte dell'universo oggi osservata risulterebbe necessariamente piatta.

Oltre a risolvere il problema della piattezza, il modello dell'universo inflazionario ottiene altri risultati decisamente positivi. In particolare, è l'unico modello che colleghi in modo coerente le condizioni iniziali che hanno causato l'espansione dell'universo alle leggi della microfisica. Il modello inflazionario permette altresì di calcolare a partire da principi primi alcuni parametri il valore dei quali si doveva in precedenza postulare o dedurre. Per esempio, il modello prevede che la forma dello spettro delle fluttuazioni di densità primordiali (la relazione funzionale tra l'ampiezza delle fluttuazioni e la loro lunghezza di scala) sia proprio quella che era stata proposta in precedenza sulla base di considerazioni fenomenologiche. Il fatto che una grandissima parte dei cosmologi accetti le previsioni del modello dell'universo inflazionario è indicativo dell'influsso profondo che la teoria delle particelle elementari esercita oggi sulla cosmologia moderna.

Proprio risolvendo il problema della piattezza il modello inflazionario rende la questione della materia oscura più pressante. Se l'universo è piatto, allora la maggior parte della sua massa probabilmente non è costituita da materia comune e la maggior parte di questa materia inconsueta non è stata ancora rilevata in alcun modo, neppure indirettamente.

Di che cosa potrebbe essere costituita questa materia esotica non rilevata? Una delle prime proposte era che si trattasse di neutrini. Queste particelle, la cui esistenza era stata postulata per la prima volta al fine di risolvere alcuni problemi relativi alla conservazione dell'energia e della quantità di moto nel decadimento dei nuclei radioattivi, interagiscono in maniera molto debole con la materia comune e risultano quindi estremamente difficili da rilevare. Cionostante, fino a oggi sono stati scoperti sperimentalmente tre tipi di neutrini, quello elettronico, quello muonico e quello tauonico. L'ipotesi originale prevedeva che i neutrini fossero privi di massa, ma non esiste alcuna ragione teo-

rica per escludere che ne abbiano una diversa da zero; tuttavia si sono stabiliti limiti sperimentali molto stretti sul suo massimo valore possibile, che risulta davvero piccolissimo: il vincolo più forte è quello sul neutrino elettronico, che deve avere una massa inferiore a circa un decimillesimo di quella dell'elettrone.

Come possibili costituenti della massa oscura, i neutrini hanno due grandi vantaggi sugli altri aspiranti al ruolo: innanzitutto si sa con certezza che esistono e, in secondo luogo, i calcoli che hanno avuto tanto successo nella descrizione della nucleosintesi primordiale inducono a ritenere anche che i neutrini leggeri debbano essere abbondanti nell'universo odierno. Quando ebbe inizio la nucleosintesi del big bang, infatti, a temperature superiori a 10^{10} kelvin, i neutrini leggeri rimanevano in equilibrio termico con la materia grazie all'interazione debole, e quindi erano altrettanto abbondanti dei fotoni. Così, come hanno valutato per primi R. Cowsik dell'Istituto Tata di ricerche fondamentali indiano e J. McLelland dell'Università di Melbourne, se i neutrini presentano oggi all'incirca la stessa densità dei fotoni che costituiscono la radiazione di fondo e se posseggono una massa compresa tra un decimillesimo e un centomillesimo di quella dell'elettrone, essi potrebbero dar luogo a una massa complessiva sufficiente a chiudere l'universo. (Questa valutazione ha ricevuto successivamente una conferma mediante calcoli più particolareggiati.)

Questa osservazione ha acquisito una notevole importanza quando V.A. Lubimov e collaboratori dell'Istituto di fisica teorica e sperimentale di Mosca hanno annunciato di aver trovato indicazioni del fatto che il neutrino elettronico ha una massa compresa in quell'intervallo. Sulla base di questo risultato sembrava che i neutrini fossero i candidati ideali al ruolo di massa dominante dell'universo. Dopo di allora, però, la plausibilità dell'identificazione della massa oscura con i neutrini leggeri si è notevolmente ridotta. In primo luogo esistono molte perplessità sperimentali non risolte riguardo al risultato del ricercatore sovietico; in effetti la recente scoperta di un gruppo di ricercatori dell'Istituto svizzero per le ricerche nucleari di Villigen (Zurigo) sembra contraddirlo. Inoltre una vasta mole di studi da parte degli astrofisici ha dimostrato che i quadri teorici di un universo in cui i neutrini predominino non sono tanto compatibili con l'osservazione quanto sembrava in precedenza.

La prima indicazione teorica in questo caso è giunta nel 1979 dalle ricerche di Scott D. Tremaine e James E. Gunn, che allora lavoravano entrambi al Caltech. Essi hanno notato, per ragioni basate tra l'altro sul principio di esclusione di Pauli, come i neutrini con massa compresa nell'intervallo in discussione non avrebbero potuto condensare abbastanza da costituire la materia oscura su scale molto più piccole di quelle delle galassie. In seguito si è riusciti a dimostrare in maniera convincente l'esistenza della materia oscura su queste scale mediante l'osservazione di galassie sferoidali nane.

Questo risultato non preclude la possibilità che i neutrini costituiscano la materia oscura su scale più grandi. Tuttavia questa ipotesi sembra incompatibile con alcuni nuovi studi teorici di notevole peso che descrivono le prime fasi dell'evoluzione dell'universo, studi che hanno dimostrato come il tipo di aggregazione su grande scala per gravitazione (cioè l'aggregazione delle galassie in ammassi e di questi in superammassi) che è probabile in un universo nel quale predomini no i neutrini non sembra simile a quello che si osserva in realtà.

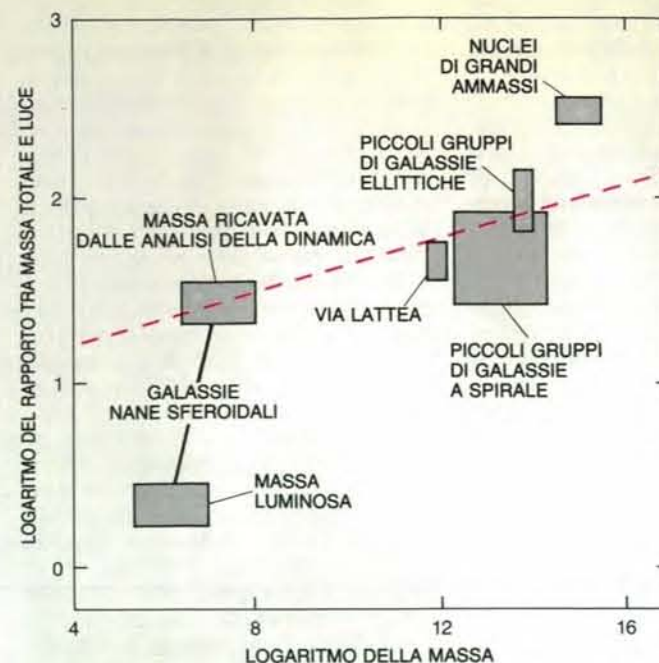
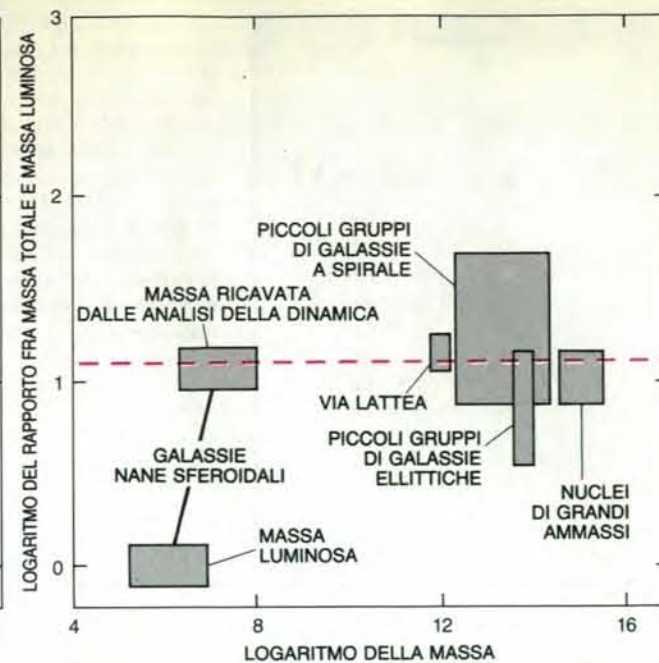
In un universo a dominante neutrinica, infatti, le prime strutture a formarsi non si troverebbero sulla scala di dimensioni delle galassie, ma su quella degli ammassi di galassie o persino di superammassi (gli ammassi di ammassi di galassie). Al contrario della materia comune, i neutrini nelle prime fasi dell'universo non erano accoppiati con la radiazione elettromagnetica che ne avrebbe sì ostacolato la condensazione, ma anche in questa situazione non riuscirono a raggrupparsi in maniera significativa perché, data la loro leggerezza, si muovevano a velocità relativistiche, e gli oggetti relativistici non vengono legati dalla gravità tranne che a corpi estremamente densi come per esempio i buchi neri.

Nel corso dell'espansione dell'universo i neutrini si raffreddarono fino a rallentare e a diventare non relativistici. Contemporaneamente la radiazione di fondo continuava a raffreddarsi fino a raggiungere energie medie inferiori a quelle dei neutrini non relativistici. I neutrini con una massa compresa nell'intervallo opportuno per permettere la chiusura dell'universo sarebbero diventati non relativistici e avrebbero cominciato a rappresentare la componente principale della densità di energia dell'universo poco prima del momento nel quale la materia comune si disaccoppiò dalla radiazione elettromagnetica. I calcoli analitici dimostrano che solo dopo questo momento i neutrini ebbero modo di aggregarsi sotto l'effetto della forza gravitazionale: in qualsiasi epoca precedente le fluttuazioni su scale maggiori di quella dell'orizzonte si sarebbero disperse perché i neutrini, essendo relativistici, non sarebbero stati legati alle regioni dense.

Così la prima scala alla quale fu possibile lo sviluppo di fluttuazioni in un universo a predominio neutrinico è quella dell'orizzonte degli eventi nel momento in cui divenne possibile per i neutrini condensare per gravità. Questa scala di distanze corrisponde all'estensione dei superammassi e non a quella delle galassie. Subito dopo il disaccoppiamento, la materia comune sarebbe stata attirata nelle buche di potenziale gravitazionale causate dagli aggregati di neutrini, formazioni delle dimensioni di un superammasso che potrebbero essersi suddivise in galassie successivamente.

Questo scenario di un universo in cui predominano i neutrini presenta molti elementi interessanti: esso avrebbe portato a un sistema di superammassi filamentosi e grandi «vuoti» (regioni prive di materia) simili alle strutture identificate durante i rilevamenti diretti a studiare su grande scala l'organizzazione in ammassi (si veda l'articolo *Strutture grandissime nell'universo* di Jack O. Burns in «Le Scienze» n. 217, settembre 1986). Inoltre la possibilità che le formazioni di neutrini legate dalla gravitazione cominciassero a formarsi prima dei sistemi composti da materia comune indica che le fluttuazioni di densità dell'universo avrebbero potuto essere abbastanza piccole da risultare marginalmente coerenti con le misurazioni dell'isotropia della radiazione di fondo.

Queste caratteristiche interessanti hanno indotto Carlos S. Frenk dell'Università di Cambridge, Simon D.M. White dell'Università dell'Arizona e Marc Davis dell'Università della California a Berkeley, e indipendentemente Joan Centrella della Drexel University e Adrian L. Melott dell'Università di Chicago, a sviluppare modelli numerici per studiare i particolari dell'aggregazione sotto l'effetto della gravità in un universo a dominante neutrinica. Quando hanno tentato di riprodurre il tipo di organizzazione in ammassi osservato in realtà, questi ricercatori hanno incontrato gravi problemi scoprendo, in sostanza, che in un universo dominato dai neutrini la suddivisione degli ammassi in galassie e la formazione di queste ultime avrebbero dovuto avvenire in un periodo relativamente vicino a noi (quando l'universo aveva almeno metà della propria età attuale) perché il livello di strutturazione degli ammassi corrispondesse a quello osservato. Questa conclusione è difficile da conciliare con l'esistenza di strutture del tipo dei quasar, che sono assai più densi del fondo e si sono formati in epoche molto precedenti.

Da un punto di vista generale il grande problema della cosmologia caratterizzata dal predomino dei neutrini è che, per permettere alle galassie di formarsi entro l'epoca attuale, le strutture di scala molto maggiore dovrebbero essere assai meno diffuse di quanto non si osservi in realtà, perché le strutture alla scala delle galassie e a quella dei superammassi si sarebbero formate contemporaneamente. L'organizzazione ben determinata in ammassi di grande scala creerebbe problemi anche al tentativo di uguagliare la velocità casuale prevista delle galassie all'interno degli ammassi con quella osservata. Per queste e altre ragioni l'ipotesi di un universo nel quale predominino i neutrini sembra oggi inaccettabile.

Le strutture di scala diversa hanno un diverso rapporto tra massa e luminosità, ma hanno circa lo stesso rapporto tra massa totale e massa luminosa. Dato che le strutture più grandi tendono ad avere un rapporto tra massa e luminosità più elevato (*a sinistra*), potrebbe sembrare che quelle di dimensioni massime contengano massa sufficiente a rendere l'universo piatto. Il rapporto fra massa totale e massa luminosa (un parametro più significativo dal punto di vista fisico) sembra però costante (*a destra*); ciò indica come le strutture più grandi non contengano in proporzione più massa di quelle piccole. La differenza tra i due grafici è dovuta in gran parte al gas caldo (che emette nella regione X dello spettro, ma non in quella visibile), il quale si trova soprattutto nei sistemi più grandi. I grafici si basano su studi condotti da George R. Blumenthal, Joel R. Primack e Sandra M. Faber dell'Università della California a Santa Cruz e da Martin J. Rees dell'Università di Cambridge.

Pare chiaro, a questo punto, quale sia la via d'uscita dai problemi dei modelli a neutrini: individuare dei modelli nei quali le galassie si possano formare in netto anticipo rispetto alle strutture più grandi. Ciò fa profilare la necessità di quella che ha assunto il nome di materia oscura fredda, una massa non emettitrice di luce così fredda (cioè in moto così lento) da risultare non relativistica ben prima dei neutrini e quindi da essere in grado di formare ammassi sotto l'azione della gravità molto prima.

Il momento nel quale un certo tipo di particelle diventa non relativistico è un fattore cruciale per stabilire le dimensioni delle strutture che quel tipo di particelle potrà formare. Prima che le particelle diventino non relativistiche le strutture di scala più piccola di quella dell'orizzonte degli eventi si disgregherebbero. Quindi, per permettere alle galassie di formarsi prima delle strutture più grandi, sarebbe necessario che la materia oscura fredda fosse stata non relativistica nel momento in cui l'orizzonte, crescendo, ha raggiunto la dimensione caratteristica delle galassie.

Fin da quando sono diventati chiari i problemi relativi alle teorie a dominante neutrinica, sono state dedicate notevoli energie all'analisi di una cosmologia nella quale predomini la materia oscura fredda, e quasi tutti i risultati sono stati positivi. Infatti, dal momento che le fluttuazioni di densità si possono sviluppare prima, non vi è bisogno che quelle iniziali siano tanto grandi, cosicché si elimina qualsiasi conflitto con l'isotropia osservata della radiazione di fondo. Inoltre la possibilità che la materia oscura fredda si sia addensata e si addensi su scale più piccole di quelle accessibili ai neutrini fa pensare che tale materia possa spiegare la massa in eccesso anche in strutture piccole come le galassie nane.

Gli studi analitici e numerici condotti in dettaglio sono estremamente incoraggianti. Si è dimostrato, per esempio, che la presenza di materia oscura fredda nelle prime fasi di evoluzione dell'universo potrebbe spiegare fin nei particolari la forma e la struttura di molti tipi di galassie. Da un punto di vista più generale Frenk e George Efstathiou di Cambridge, insieme a Davis e White, hanno dimostrato mediante tecniche numeriche che la formazione di ammassi di grande scala in un universo dove predomini la materia oscura fredda può riprodurre bene la maggior parte delle caratteristiche osservate nella struttura effettiva degli ammassi.

Rimane ancora almeno un ostacolo che sembra impedire l'accordo completo fra teoria e osservazione nel caso che l'universo sia esattamente chiuso e dominato da materia oscura fredda: dov'è questa materia? Pare che riesca a condensarsi facilmente a scale galattiche ma, come ho detto precedentemente, non esistono indicazioni della presenza a queste scale di una densità critica. Si può risolvere il problema supponendo che le stesse galassie non siano indicatori validi della collocazione della maggior parte delle grandi concentrazioni di massa, cioè che la materia oscura fredda occupi in prevalenza regioni senza alcun legame con la posizione di questi sistemi luminosi. Potrebbe darsi benissimo che le galassie rappresentino eventi rari dal punto di vista statistico e che la maggior parte della massa dell'universo non si sia mai condensata in galassie. Se così fosse, l'esame dell'organizzazione in ammassi delle galassie darebbe un valore alterato della densità effettiva di massa dell'universo. Le conseguenze di questa proposta sono state oggetto di studi particolareggiati, dai quali sembra che essa conduca a scenari in buon accordo con quasi tutti gli aspetti dell'aggregazione in ammassi osservata (a parte alcune eccezioni notevoli). Inoltre gli studi che compiono attualmente Frenk e collaboratori inducono a ritenere che gli scenari nei quali le galassie risultano rare in senso statistico potrebbero derivare dai modelli della formazione di ammassi sotto l'azione della gravitazione in maniera più naturale di quanto non si supponesse in precedenza.

L'ipotesi della materia oscura fredda ha stretto un saldo legame fra la fisica delle particelle e la cosmologia. Proprio mentre i cosmologi decidevano

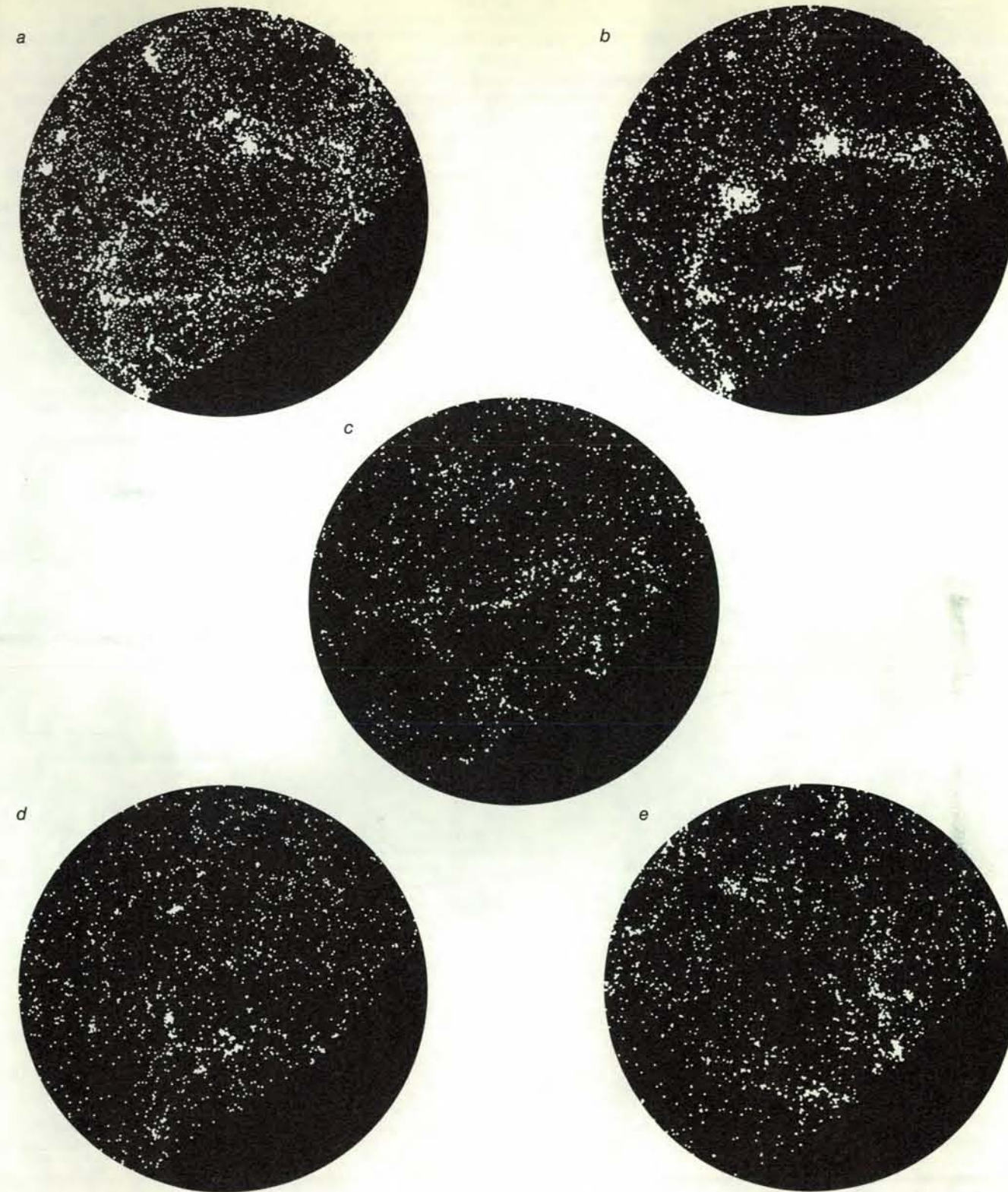
che occorreva qualche genere di materia oscura fredda, i fisici delle alte energie stavano prospettando in via del tutto indipendente la possibilità che, nel contesto di varie teorie di unificazione, esistessero nuove particelle esotiche. Si è dato il caso che numerose particelle di cui si è ipotizzata l'esistenza per colmare le lacune teoriche della fisica delle alte energie possano fungere in modo assai naturale anche da materia oscura fredda per i cosmologi. Il loro svantaggio è che non sono ancora state osservate: al contrario dei neutrini, in questo momento si tratta solo di costruzioni teoriche. Eppure posseggono il grande pregio di aver visto la propria esistenza proposta in un contesto indipendente dalla cosmologia: sono state ipotizzate come soluzione ad alcuni problemi disparati della teoria delle particelle elementari, ma ciascuna di esse, per motivi completamente diversi, potrebbe fungere da costituente della materia oscura fredda.

Uno degli aspiranti al ruolo di costituente della materia oscura più quotati oggi sul mercato è l'assione, una particella la cui esistenza deriva in maniera naturale da un'impostazione teorica elaborata per spiegare un nesso speciale che nella teoria delle interazioni forti tra quark collega le due forme di simmetria dette coniugazione di carica e parità.

Un'interazione si dice simmetrica per coniugazione di carica se mantiene lo stesso «aspetto» nel caso che a ogni particella sia sostituita la sua antiparticella (che ha la carica opposta). Un'interazione si dice simmetrica per parità se non muta aspetto per riflessione speculare. Le interazioni governate dalla forza nucleare forte (la forza che lega i quark in protoni e in neutroni) sembrano simmetriche a un grado elevatissimo per una combinazione particolare della coniugazione di carica e della parità: hanno un aspetto pressoché identico se si sostituiscono tutte le particelle con le antiparticelle corrispondenti e si sottopone a riflessione speculare tutta l'interazione risultante. Dal punto di vista teorico, questa combinazione particolare di simmetrie non deve risultare valida necessariamente: le equazioni che governano le interazioni forti contengono diversi termini che in linea di principio potrebbero violare in maniera plateale questo tipo di combinazione.

Nel 1977 Roberto D. Peccei e Helen R. Quinn, che allora lavoravano alla Stanford University, hanno suggerito un modo per spiegare come mai quella combinazione di simmetrie risulta rispettata con tale precisione. La soluzione consisteva nell'introdurre una nuova forma di simmetria, un rapporto tra le forme di forze fondamentali diverse che risulta evidente a energie sufficientemente elevate, ma si spezza a basse energie. In seguito Frank Wilczek, che oggi si trova all'Università della California a Santa Barbara, e Steven Weinberg dell'Università del Texas ad Austin hanno sottolineato come il fatto che la simmetria di Peccei-Quinn si rompa indichi l'esistenza di una nuova particella molto leggera. Questa particella è l'assione. Molti studi teorici, negli ultimi tempi, hanno perfezionato il modello originale aumentando la temperatura alla quale si prevede che tale simmetria si spezzi. Una delle grandi sorprese che ne sono derivate è quella che, dipendendo l'esistenza degli assioni da una rottura di simmetria, si potrebbe formare nell'universo un «campo di fondo» di assioni, più o meno come avverrebbe per un campo elettrico se l'universo non fosse simmetrico rispetto alla carica (cioè se non contenesse cariche positive e negative in ugual numero). Gli assioni sono anch'essi molto leggeri, ma i calcoli indicano che il campo di fondo nel proprio complesso potrebbe agglomerarsi in modo analogo a quello delle particelle non relativistiche, ciò che lo rende un aspirante ideale al ruolo di costituente della materia oscura.

Un altro aspirante deriva dal quadro teorico che prende il nome di «supersimmetria». In questa teoria per ogni particella oggi conosciuta esiste un «partner supersimmetrico», cioè una particella identica sotto molti aspetti tranne, fra gli altri, lo spin. Queste particelle non sono ancora state osservate in laboratorio e quindi devono essere di massa elevata. Modelli semplici inducono a ritenere che le particelle supersimmetriche possano comportarsi, nelle interazioni con la materia comune, in modo simile a quello di eventuali neutrini molto pesanti. L'aspirante più quotato al ruolo di costituente della materia oscura tra le particelle supersimmetriche è la controparte supersimmetrica del fotone, ossia il fotino. I calcoli compiuti da chi scrive e da altri indicano che i fotini nell'intervallo di masse compreso tra una e 50 volte quella del protone potrebbero possedere in maniera naturale un'abbondanza cosmica sufficiente a chiudere l'universo oggi. Anche se negli ultimi tempi questa proposta ha dato luogo a un notevole entusiasmo, è bene osservare come i modelli che prevedono l'esistenza di fotini portino ad altre previsioni cosmologiche che risultano difficili da conciliare con le osservazioni.

Un ultimo partecipante alla sfilata degli aspiranti alla costituzione della materia oscura fredda non è neppure una particella, ma una struttura che prende il nome di «corda cosmica». Le corde cosmiche sono difetti topologici estesi che potrebbero essere derivati dalla rottura di una simmetria nelle prime fasi della storia dell'universo, oggetti che assumerebbero la forma di tubi molto lunghi e sottili con densità di energia costante e assai elevata dipanantisi per tutto l'universo. Si è dedicato un grande impegno a far vedere come le corde cosmiche si sarebbero potute evolvere in modo tale che la loro densità di energia complessi-

Le simulazioni numeriche degli scenari di formazione di una struttura di grande scala nell'universo inducono a ritenere che la materia oscura non sia costituita da neutrini leggeri. Può darsi che sia formata da un tipo di materia oscura fredda, particelle che nelle prime fasi della evoluzione dell'universo erano non relativistiche e che quindi erano in grado di aggregarsi gravitazionalmente prima dei neutrini. I disegni illustrano l'organizzazione delle galassie in ammassi in ogni scenario, ipotizzando condizioni iniziali plausibili; i punti rappresentano le galassie. In un universo pieno di neutrini leggeri in quantità sufficiente a spiegare tutta la materia oscura (*a*, *b*), l'organizzazione in ammassi su grande scala sarebbe più definita di quanto non indichino le osservazioni (*c*). Un universo in cui al momento della formazione predominasse la materia oscura fredda (*d*, *e*) riproduce molto meglio i risultati delle osservazioni. Negli scenari a materia oscura illustrati l'universo contiene solo il 20 per cento circa della materia necessaria a renderlo piatto. Le simulazioni nelle quali l'universo contiene abbastanza massa da essere piatto non riproducono bene le osservazioni se la distribuzione delle galassie è indicativa di quella della massa che predomina nell'universo. Se invece le galassie non riflettono la distribuzione della materia che predomina, le simulazioni che utilizzano la materia oscura fredda concordano molto bene con la struttura su grande scala osservata. Le simulazioni sono state compiute da Marc Davis dell'Università della California a Berkeley, George Efstathiou e Carlos S. Frenk dell'Università di Cambridge e Simon D. M. White dell'Università dell'Arizona.

va attuale sarebbe sufficiente a chiudere l'universo. In un universo dove predominino materia oscura fredda e corde il meccanismo della formazione delle galassie potrebbe essere molto diverso da quello valido nei modelli classici basati sulla materia oscura fredda, ma la struttura degli ammassi risultante potrebbe comunque corrispondere alla struttura osservata.

Quello che rende tanto interessanti tutti questi aspiranti a costituire la materia oscura è, in questo momento, l'idea che ciascuno di essi potrebbe essere rilevato, direttamente o indirettamente, nel futuro prossimo. Sono infatti possibili alcuni esperimenti in grado di escludere o, più significativo ancora, di confermare diverse di queste ipotesi. Un risultato positivo in uno qualunque di questi esperimenti fornirebbe informazioni preziosissime sull'evoluzione della struttura su grande scala dell'universo e sulla struttura fondamentale della materia, e inoltre potrebbe mettere a disposizione un meccanismo insostituibile per studiare la successione di eventi che ebbe luogo durante i primi pochi secondi del big bang stesso.

Pierre Sikivie dell'Università della Florida è stato il primo a mettere in evidenza come si potrebbero rilevare gli assioni cosmici, che pure interagiscono con l'altra materia in modo estremamente debole, all'interno di una cavità risonante (una cavità nella quale risuona la radiazione elettromagnetica a certe frequenze nella regione delle microonde). Un campo di fondo di assioni in oscillazione concorde potrebbe produrre una radiazione elettromagnetica suscettibile di essere rilevata, in linea di principio, in un ricevitore per microonde. Wilczek, John Moody dell'Università della California a Santa Barbara, Donald E. Morris del Lawrence Berkeley Laboratory e chi scrive hanno esaminato

1
NUCLEO DI SILICIO
PARTICELLA DI MATERIA OSCURA IN ARRIVO
SENSORI DI ONDE SONORE
SILICIO PURISSIMO ULTRAFREDDO

2
SENSORI DI ONDE SONORE
SILICIO
FONONI (ONDE SONORE)

3

3a
ALLUMINIO SUPERCONDUTTORE
OSSIDO D'ALLUMINIO
FONONE
ELETTRONE
SILICIO ULTRAFREDDO
COPPIA DI COOPER

3b
CORRENTE →
TEMPO (MICROSECONDI) →

4
SENSORE DI TEMPERATURA

4a
MATERIALE TERMOSENSIBILE
SILICIO

4b
TEMPERATURA →
TEMPO (MILLISECONDI) →

I dispositivi per il rilevamento delle particelle di materia oscura potrebbero essere costituiti da cristalli di silicio molto puri raffreddati entro un grado dallo zero assoluto. Cristalli di questo genere potrebbero reagire in maniera misurabile a depositi di energia estremamente piccoli. In un tipo di rivelatori possibile, quando una particella di materia oscura incide sullo strumento, essa diffonde il nucleo di un atomo di silicio (*1*) e fa propagare per tutto il materiale un insieme di fononi, cioè di onde sonore (*2*). I fononi che raggiungono la superficie del silicio presenteranno una configurazione peculiare (*3*), che dipenderà dalla posizione e dall'intensità della collisione originaria. In una delle configurazioni il rivelatore potrebbe individuare i singoli fononi all'interno della struttura complessiva, quando collidono con la superficie del cristallo. Per far ciò si potrebbe ricoprire il silicio con strisce formate da due strati di alluminio superconduttore inframezzati da uno di ossido di alluminio (*3a*). Nei superconduttori gli elettroni sono legati in coppie, le coppie di Cooper. Un fonone incidente potrebbe scinderne una, e se gli strati di alluminio vengono mantenuti a tensioni elettriche diverse gli elettroni così liberati potrebbero passare per effetto tunnel dall'uno strato all'altro, dando origine a una corrente elettrica (*3b*). In alternativa si potrebbe misurare l'aumento della temperatura del silicio dopo la dissipazione dei fononi di energia relativamente elevata iniziali in un fondo uniforme di vibrazioni termiche casuali (*4*). In questo caso la componente attiva del rivelatore sarebbe costituita da una pellicola sottile di un materiale la cui resistenza elettrica aumenta bruscamente con la temperatura (*4a*). Una variazione di temperatura complessiva del campione (*4b*) si potrebbe ricavare dalla variazione di resistenza.

questo schema di rilevamento nei particolari, proponendo varie migliorie e alternative. Sembra che la sensibilità necessaria per rilevare gli assioni cosmici sia vicina al limite della tecnologia attuale, la quale però sta progredendo in modo molto rapido.

Altri aspiranti al ruolo di costituenti della materia oscura, dotati di grande massa come il fotino, potrebbero essere rilevati in vari modi. Negli ultimi tempi chi scrive, come altri ricercatori indipendentemente, ha avanzato l'ipotesi che queste particelle pesanti possano restare imprigionate nel nucleo del Sole o della Terra, dove si accumulerebbero. In tali regioni, come hanno indicato i calcoli successivi, potrebbero collidere con le proprie antiparticelle (anch'esse passibili di cattura) in reazioni di annichilazione che potrebbero a loro volta produrre neutrini leggeri, i quali forse sfuggirebbero dal nucleo della Terra o del Sole e si potrebbero captare con grandi rivelatori sotterranei. Il grado di certezza con il quale un flusso di neutrini leggeri di questo tipo non è stato ancora osservato stabilisce un limite superiore alla massa e alla densità di questo tipo di aspiranti.

Negli ultimi tempi si è messo in evidenza come le particelle pesanti della materia oscura si potrebbero anche rilevare direttamente con dispositivi sensibili a depositi molto piccoli di energia in volumi di materiale grandissimi, strumenti dei quali è stata proposta tutta una serie. Uno di questi dispositivi, ipotizzato da Blas Cabrera di Stanford, da Wilczek e da chi scrive, è progettato in modo da misurare un piccolo aumento della temperatura di un grande campione di silicio o di un altro materiale cristallino puro freddissimo. L'aumento di temperatura avverrebbe quando le onde sonore prodotte dalle particelle della materia oscura incidenti diffondessero in modo casuale. Gli studi di Cabrera, Barbara Neuhauser e Jeffrey C. Martoff di Stanford inducono a ritenere che forse sarebbe possibile rilevare direttamente le stesse onde sonore (*si veda l'illustrazione nella pagina a fronte*).

Persino le corde cosmiche potrebbero risultare accessibili al rilevamento tra breve, sia grazie agli effetti gravitazionali diretti che avrebbero sulla luce proveniente dai quasar lontani e sul fondo a microonde (le concentrazioni di energia con densità elevata come quella delle corde cosmiche dovrebbero creare campi gravitazionali che devierebbero la luce in maniera non trascurabile), sia indirettamente, mediante la misurazione delle onde gravitazionali o di altra radiazione che dovrebbero emettere mentre si evolvono.

La soluzione del problema della materia oscura potrebbe avere effetti profondi su molte aree della fisica e dell'astronomia. Sono in gioco i concetti fondamentali tanto della cosmologia quanto della fisica delle particelle elementari, ed è significativo che ognuna delle due discipline abbia svolto – spesso stimolando nell'altro campo un vivace dibattito – un ruolo importante nell'evoluzione simbiotica di questa area della ricerca.

È importante tuttavia rendersi conto del fatto che la cosmologia sta muovendo sotto molti aspetti i primi passi: i dati sperimentali e osservativi disponibili per il lavoro dei teorici sono relativamente scarsi, tanto da rendere possibile che avvengano grandi colpi di scena e che gran parte delle conoscenze ritenute oggi assodate si riveli errata. Validi esempi di questa situazione vengono da diversi risultati nuovi prospettatisi durante la stesura di questo articolo, risultati di tale portata che uno qualsiasi di essi potrebbe avere ripercussioni profonde su tutta la disciplina.

Un nuovo risultato osservativo viene dall'analisi preliminare di un rilevamento del cielo in profondità che stanno compiendo Margaret J. Geller, John P. Huchra e collaboratori allo Harvard-Smithsonian Astrophysical Observatory. Sembra che le galassie vicine si ammassino in superfici dall'aspetto pellicolare che avvolgono vuoti quasi sferici, una struttura simile a quella delle bolle di sapone o della schiuma. Questa importante osservazione, la conferma della quale porterebbe a una revisione completa dell'immagine che i cosmologi hanno della struttura su grande scala, fa pensare che altre forze, oltre alla gravitazione, contribuiscano forse a stabilire la struttura su grande scala dell'universo odierno.

Un altro nuovo sviluppo è quello che riguarda gli studi compiuti indipendentemente da Tremaine (che attualmente lavora al Canadian Institute of Theoretical Astrophysics) e da J. Anthony Tyson degli AT&T Bell Laboratories, ricerche le quali inducono a ritenere che la curva di rotazione delle galassie non resti piatta a distanze indefinitamente grandi, ma scenda invece per raggi superiori a circa 30 chiloparsec. Questo risultato comporta che quanto costituisce la materia oscura, di qualunque cosa si tratti, interagisca con la materia comune in maniera più intensa di quanto ci si aspetterebbe dalla materia oscura fredda.

Infine, alcuni dati emersi negli ultimi tempi sul moto rispetto al fondo a microonde della materia contenuta in regioni di scala molto grande indicano che queste regioni si spostano, solidali, con una velocità di deriva estremamente elevata. Nessuna delle teorie attuali sulle strutture di grande scala è in grado di spiegare ciò che perlomeno sembra di avere osservato. Altre nuove misurazioni come queste, oltre alla possibilità di rilevare la stessa materia oscura, possono rivoluzionare in breve tempo l'immagine dell'universo che si dà per assodata.

La struttura funzionale della retina

Nuove tecniche permettono di studiare la disposizione e le reciproche connessioni di intere popolazioni di cellule e di concludere che decine di tipi cellulari hanno ruoli specifici nella codificazione del mondo visivo

di Richard H. Masland

La retina codifica il mondo visivo, cioè trasforma le immagini ottiche in treni di impulsi nervosi, che poi sono convogliati lungo il nervo ottico e trasmessi al cervello. Il cervello interpreta questi segnali e li trasforma in percezione visiva, una sensazione soggettiva della forma, dei colori e dei movimenti che circondano l'osservatore. La retina, comunque, è qualcosa di più di una serie di fotocellule: questo sottile strato di tessuto nervoso, posto nella parte posteriore dell'occhio, rappresenta un avamposto del sistema nervoso centrale. I suoi circuiti di neuroni, o cellule nervose, reciprocamente connessi eseguono una specie di analisi dell'immagine; certi aspetti dello stimolo visivo non ancora elaborato vengono accentuati, mentre altri sono minimizzati.

Il tentativo di comprendere come i segnali luminosi siano trasdotti in attività nervosa e come questa attività venga a sua volta interpretata ha impegnato i neurobiologi per molti anni. In primo luogo è stata stabilita la natura generale della codificazione del mondo visivo operata dalla retina, ovvero la relazione tra segnali visivi in entrata (*input*) e in uscita (*output*). All'inizio degli anni settanta è stata raggiunta una visione schematica dei mezzi mediante i quali questa codificazione viene compiuta. Si sapeva che la retina era composta di cinque classi principali di neuroni, che sono connessi gli uni con gli altri mediante sinapsi, punti di stretta giustapposizione, dove i messaggeri chimici (neurotrasmettitori) sono liberati da un neurone per influenzare il neurone vicino. Tre delle cinque classi di neuroni retinici formano una via diretta dalla retina al cervello; si tratta dei fotorecettori (bastoncelli e coni), delle cellule bipolari e delle cellule gangliari. Le altre due classi di neuroni retinici, le cellule orizzontali e le cellule amacrine, formano vie laterali, che modificano e controllano il messaggio che deve passare lungo la via diretta.

Ancora di recente si pensava che le cinque classi di cellule retiniche sopra citate definissero esattamente gli elementi funzionali della retina e si riteneva che ciascuna classe eseguisse un solo tipo di compito; per comprendere i codici interni della retina, i neurobiologi si sarebbero dovuti limitare a decifrare le interazioni di questi elementi fondamentali, relativamente poco numerosi. Oggi, invece, si sa con certezza che la retina è costituita da un numero di elementi funzionali molto superiore a cinque. Le classi di cellule retiniche comprendono sottotipi così differenziati che il numero reale degli elementi funzionali può arrivare a cinquanta.

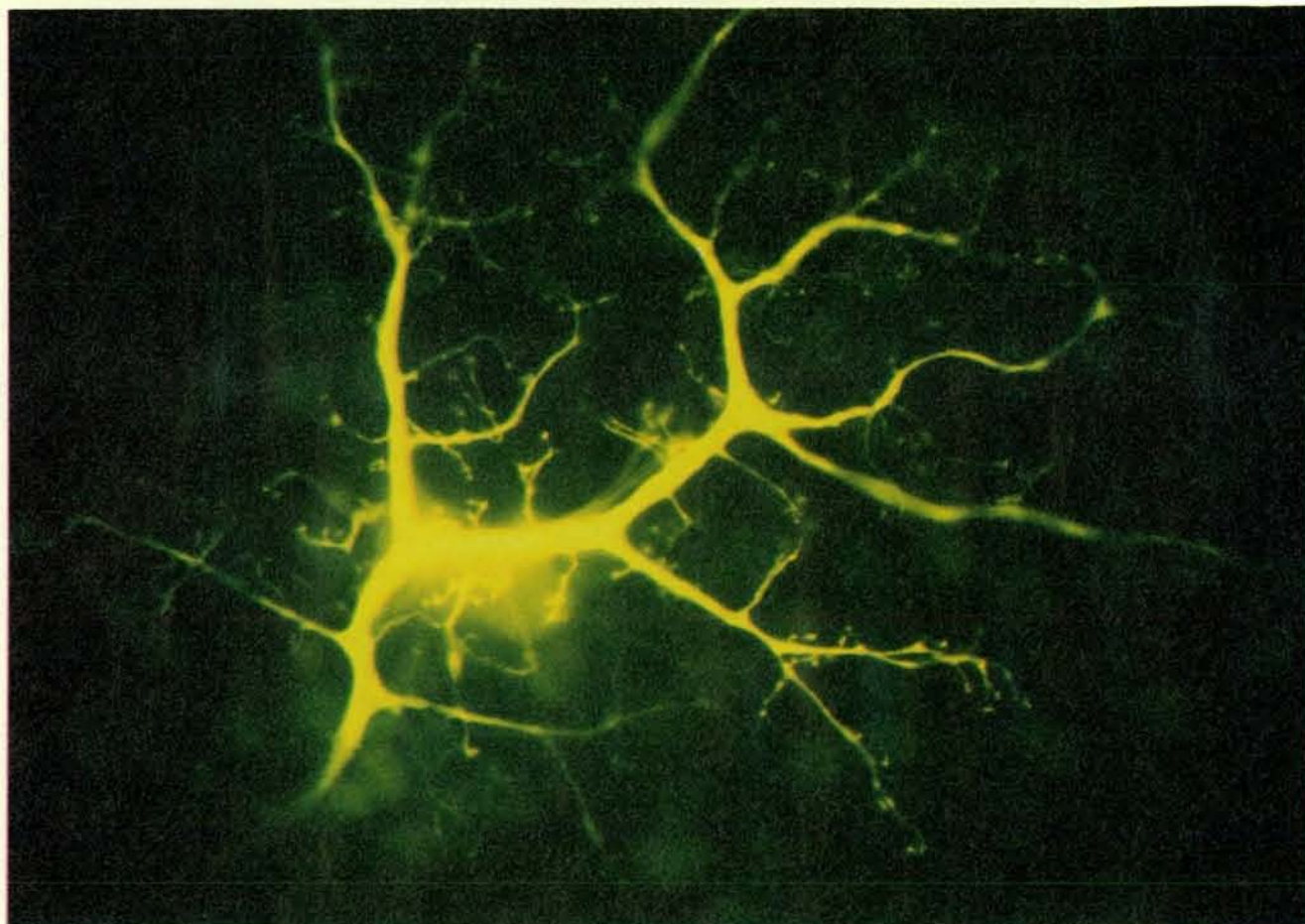
Si tratta quindi di comprendere perché sono necessari tanti tipi di cellule: un compito lungi dall'essere stato portato a termine, ma nel quale nei pochi anni trascorsi, sono stati compiuti importanti progressi. Un passo avanti significativo è stato lo sviluppo di mezzi per esaminare forma e distribuzione di intere popolazioni di cellule nervose nella retina.

La codificazione dell'informazione visiva da parte della retina dei mammiferi ha cominciato a essere compresa nel 1952, con una serie di esperimenti sulle cellule gangliari, quelle cellule i cui assoni formano il nervo ottico. I primi esperimenti, compiuti da Stephen W. Kuffler alla Johns Hopkins University Medical School, hanno cercato di rispondere alla domanda: come si modifica l'attività elettrica delle cellule gangliari in risposta alla luce? Kuffler ha registrato l'attività elettrica delle singole cellule gangliari e ha trovato che la maggior parte di queste cellule dà origine a una corrente continua di potenziali di azione, o impulsi nervosi, lungo l'assone, anche in assenza di luce. Quindi una cellula gangliare risponde alla presenza di luce incrementando in modo marcato o diminuendo la frequenza di scarica degli impulsi nervosi.

Esaminando con sottili fasci di luce la superficie della retina di gatti anestetizzati, Kuffler ne ha studiato regioni specifiche e ha così potuto misurare la risposta elettrica di singole cellule. Egli ha trovato che ciascuna cellula gangliare risponde a una precisa area del campo visivo che è il suo campo recettivo, di dimensioni variabili, ma comunque generalmente piccole. Nel gatto i centri dei più piccoli campi recettivi occupano una regione della retina che ha un diametro di circa 120 micrometri (milionesimi di millimetro). Se il gatto guarda una parete a due metri di distanza, ciascuna di queste cellule gangliari prende in considerazione un'area che ha grosso modo il diametro di un centimetro.

Le cellule che Kuffler ha studiato si possono suddividere in due popolazioni: quelle che vengono stimolate e quelle che vengono inibite quando si stimola il centro del loro campo recettivo. Il centro agisce inoltre in opposizione alla regione che lo circonda (la periferia). Se la stimolazione del centro del campo recettivo eccita la cellula, la stimolazione della periferia la inibisce. All'inverso, se il centro è inibito dalla luce, la periferia ne è eccitata. In altre parole, la periferia svolge un ruolo «antagonista», mentre la cellula attua una semplice forma di aumento di contrasto.

Subito dopo è stato identificato un tipo di cellula gangliare molto diverso: una cellula con selettività direzionale e il cui campo recettivo risponde alla direzione di uno stimolo in movimento (è risultato che essa è solo un elemento di una serie di cellule che svolgono complesse funzioni selettive nei confronti degli stimoli, ma di queste cellule è la più numerosa e meglio studiata e può per-

Nella fotografia, eseguita dall'autore, è visibile una retina di coniglio. Dal punto di vista istologico, la retina è una lamina trasparente di tessuto nervoso che riveste l'interno della parte posteriore dell'occhio e che ha uno spessore medio di circa un decimo di millimetro. La luce che l'attraversa colpisce i bastoncelli e i coni, che sono vicini alla sua superficie posteriore. Questi fotorecettori trasducono il segnale luminoso in un segnale nervoso, che viene poi trasmesso da strati successivi di neuroni retinici; esso finisce per eccitare le cellule gangliari i cui assoni (*filamenti bianchi*) contribuiscono a formare il nervo ottico. Si vedono anche due paia di vasi sanguigni che attraversano la superficie anteriore della retina. Metodi per isolare la retina vivente sono stati sviluppati da Adelbert Ames III del Massachusetts General Hospital.

In questo neurone retinico, che è una cellula orizzontale, è stato iniettato un colorante fluorescente, il giallo Lucifero, che impregna tutti i prolungamenti della cellula. Nella microfotografia, di Julie H. Sandell, si nota un'arborizzazione terminale della cellula: l'informazione fluisce lungo queste ramificazioni sempre più sottili, fino a raggiungere i bottoni terminali, che stabiliscono un contatto sinaptico con le altre cellule.

tanto servire come prototipo). Le cellule con selettività direzionale sono state scoperte nella rana da Jerome Y. Lettvin, Humberto R. Maturana, Walter H. Pitts e Warren S. McCulloch, tutti del Massachusetts Institute of Technology. Solo parecchi anni dopo, comunque, le loro esatte proprietà sono state chiarite in una serie di esperimenti, oramai classici, compiuti su conigli da Horace B. Barlow e William R. Levick dell'Università di Cambridge.

Il comportamento delle cellule con selettività direzionale varia a seconda che lo stimolo luminoso si muova o sia stazionario. Se si esplora il campo recettivo della cellula mediante uno stimolo luminoso stazionario, si osserva in tutta la sua estensione una risposta uniforme della cellula. Quando, invece, un punto luminoso viene mosso lentamente attraverso il campo, il comportamento della cellula cambia nettamente: scarica un treno continuo di potenziali d'azione quando il punto luminoso si muove in una direzione e scarica poco o niente quando il punto si muove nella direzione opposta. Questi primi risultati hanno definito le questioni fondamentali riguardanti i circuiti della retina. Che cosa definisce il campo recettivo di una cellula gangliare? Come può il sistema neuronale trasformare l'input ottico in un modo così selettivo? Per poter fornire delle risposte, è necessario considerare più in particolare le componenti interne della retina.

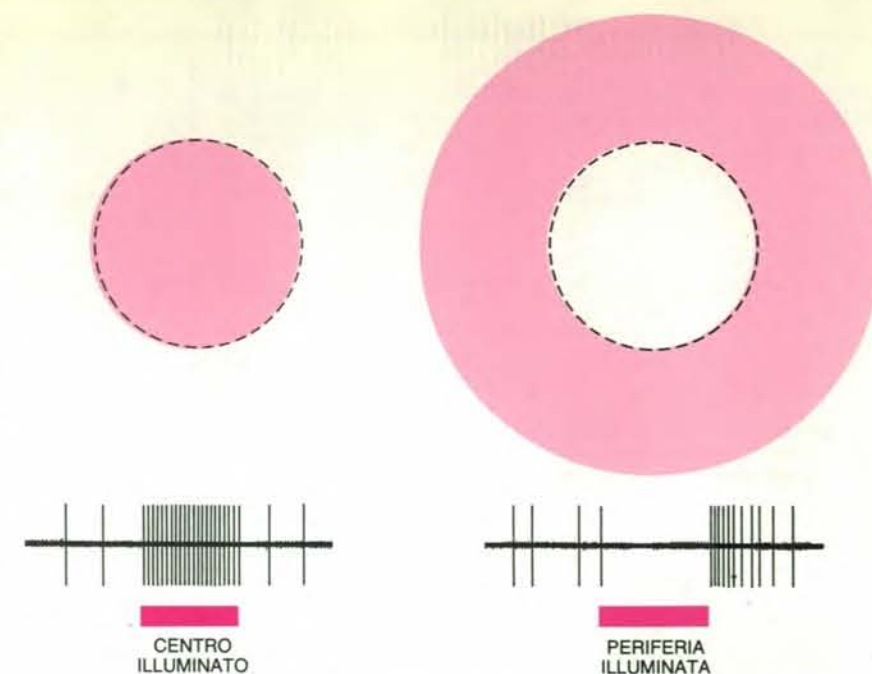
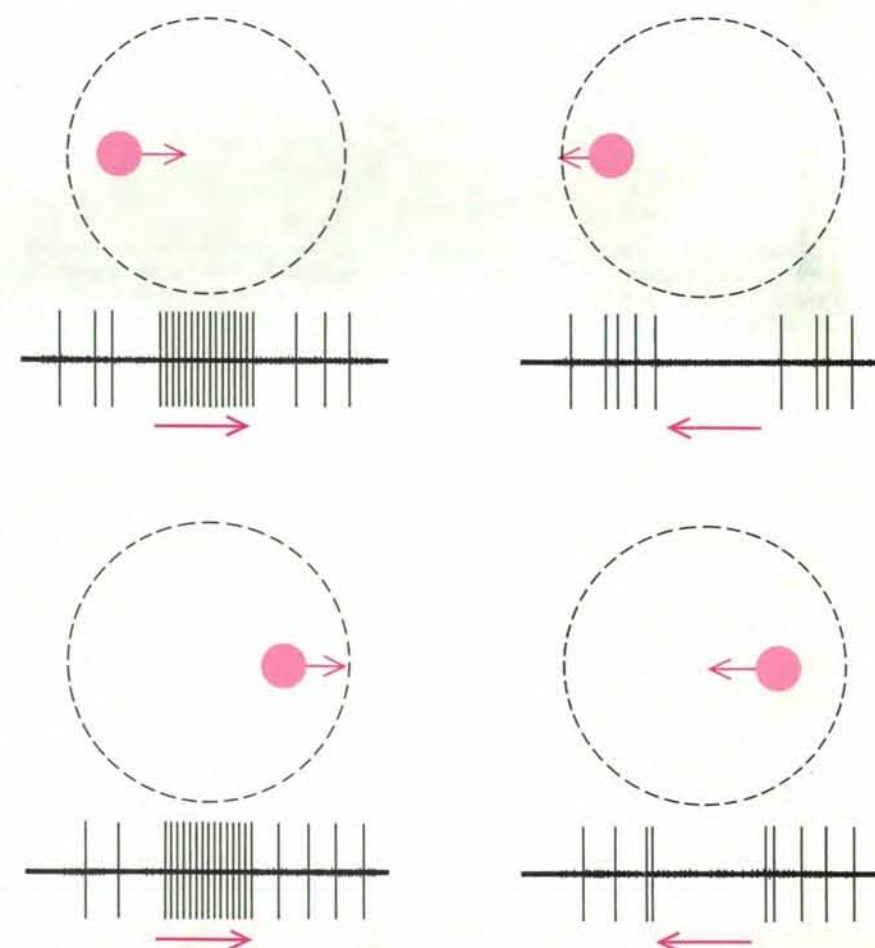
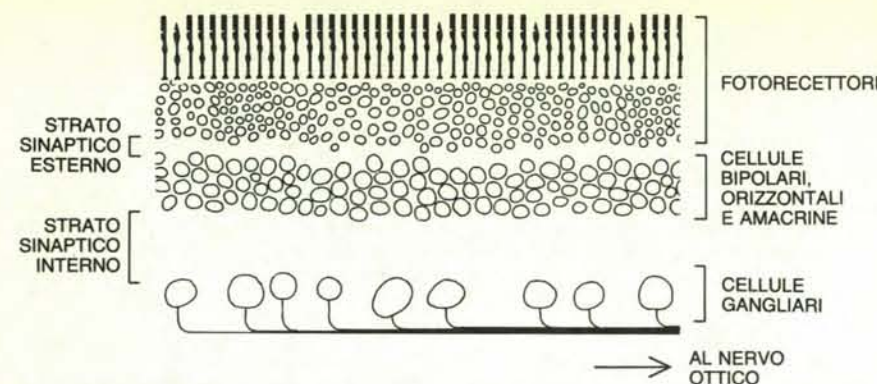
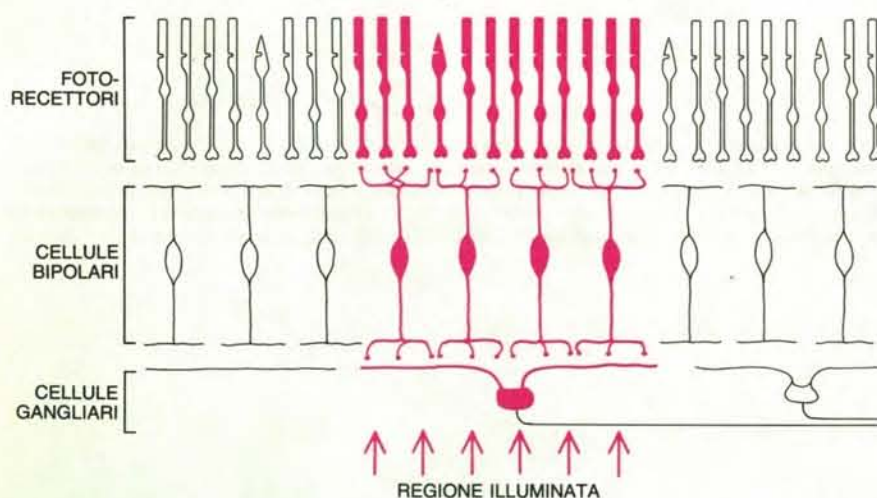
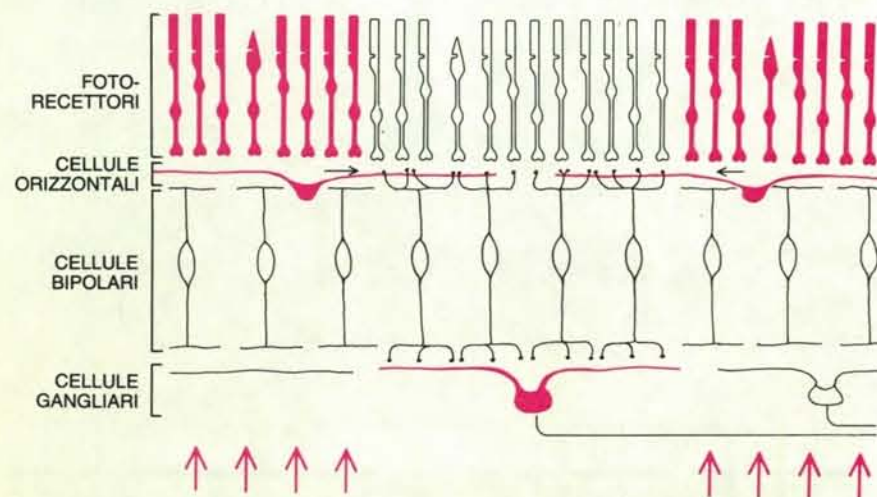
Come ho già fatto presente, la retina dispone di due serie di neuroni: quelli che stabiliscono una via diretta dalla sorgente luminosa al nervo ottico e quelli che instaurano connessioni laterali. È così subito balzato in evidenza il fatto che la presenza di una periferia antagonista, osservata in alcune cellule gangliari, trova una sua giustificazione se i neuroni della retina a conduzione laterale funzionano in opposizione ai neuroni a conduzione diretta. In altre parole, quando il centro di una cellula gangliare risponde alla luce (per via diretta), la periferia (che agisce in opposizione) potrebbe essere eccitata da stimoli trasmessi attraverso le cellule orizzontali o le cellule amacrine. Una chiara conferma nei confronti di uno dei ruoli attribuiti alle cellule amacrine si è avuta quando si è appreso che le cellule bipolari hanno periferie antagoniste. Dal momento che la cellula bipolare è il solo tipo di cellula che conduce gli stimoli dalla retina esterna (dove si trovano le cellule orizzontali) alla retina interna, non era necessario pensare ad alcun altro meccanismo per spiegare l'esistenza di periferie antagoniste.

Altri neuroni laterali, le cellule amacrine, regolano il comportamento di certe cellule gangliari (comprese quelle a selettività direzionale) che danno una risposta transitoria alla stimolazione da parte della luce. Quando un raggio luminoso colpisce la retina, queste cellule scaricano immediatamente un treno di potenziali d'azione, ma la scarica cessa quando la stimolazione luminosa si prolunga. I fotorecettori (con i bastoncelli), le cellule bipolari e le cellule orizzontali non esibiscono un comportamento transitorio a questo livello, che è invece caratteristico di molte cellule amacrine. Uno dei ruoli di queste cellule, quindi, è presumibilmente quello di intensificare le risposte transitorie di certe cellule gangliari, fra le quali ci sono quelle a selettività direzionale.

Le cellule amacrine, comunque, devono avere altri compiti oltre a questo. Esse presentano una morfologia estremamente varia: una sola retina ne può contenere fino a 30 tipi distinti. Se la loro sola funzione fosse quella di rendere transitoria la risposta di alcune cellule gangliari, ci si potrebbe ragionevolmente attendere un solo tipo di cellule amacrine, o al massimo pochi tipi.

La diversità di forma delle cellule amacrine era stata osservata fin dal 1892 da Santiago Ramón y Cajal, ma per un lungo periodo di tempo le scoperte di questo grande neuroanatomista spagnolo furono misconosciute. Tra le motivazioni addotte da coloro che sottovalutavano i suoi lavori, ve n'era una che sosteneva che le differenti forme delle cellule da lui descritte fossero semplici varianti di uno stesso tipo cellulare: le cellule sembravano diverse, ma avevano la stessa funzione. Solo alla fine degli anni sessanta, quando Berndt Ehinger dell'Università di Lund in Svezia iniziò a studiare la biochimica delle cellule amacrine, la vera diversità di queste cellule fu riconosciuta. Ehinger applicò alla retina i metodi che da poco erano stati messi a punto per identificare i neurotrasmettitori in sottili sezioni di tessuto nervoso. Egli trovò che numerosi neurotrasmettitori, già noti nel cervello, erano presenti anche nei neuroni della retina. La cosa sorprendente era che tutti questi neurotrasmettitori venivano trovati o nell'una o nell'altra cellula amacrina, e ciascun neurotrasmettitore si trovava in un determinato sottoinsieme di cellule, il che implicava che vi fossero molti distinti sottoinsiemi di questi tipi cellulari.

I risultati di Ehinger furono guardati dapprima con un certo scetticismo. I metodi erano nuovi e si poteva sostenere che avessero sovrastimato o sottostimato le popolazioni di cellule amacrine. Eppure gli studi successivi, condotti da altri ricercatori, portarono agli stessi risultati. Tra questi studi vi era una ricerca eseguita da John W. Mills e da me presso il Massachusetts General Hospital ed eseguita su cellule amacrine, che sintetizzano il neurotrasmettitore acetilcolina. Abbiamo incubato retine di coniglio isolate in presenza di colina marcata con elementi radioattivi, cioè del precursore dal quale le cellule sintetizzano l'acetilcolina. Subito dopo ci siamo posti come obiettivo l'immobilizzazione della acetilcolina per poterla poi identificare mediante autoradiografia (un metodo che localizza la radioattività in sezioni di tessuto) e determinare così quali cellule l'avessero sintetizzata. Per far questo, siamo ricorsi a un metodo di rapido congelamento: campioni di retine vive sono stati immersi direttamente in propano a −180 gradi Celsius, sono quindi stati essiccati per congelamento e preparati per l'autoradiografia con metodi che evitano di metterli a contatto con l'umidità.

Si tratta di procedimenti noiosi, ma che presentano un particolare vantag-

Il campo recettivo concentrico presenta una regione centrale in cui la stimolazione da parte della luce produce un certo effetto sull'eccitazione della cellula gangliare (registrato da un elettrodo che misura l'attività di un singolo assone del nervo ottico) e una regione periferica, in cui l'effetto è opposto (*in alto*). Nel caso di una cellula centrale, la stimolazione del centro fa aumentare la sua attività (*in basso a sinistra*), la stimolazione della periferia la fa diminuire (*in basso a destra*).

Una cellula con selettività direzionale reagisce quando un punto luminoso si muove in una direzione (verso destra nella figura) ed è inibita quando il punto si muove nella direzione opposta. Questo tipo di percezione è indipendente dal punto del campo recettivo in cui cade la luce.

I corpi cellulari dei neuroni della retina sono distribuiti in tre strati. Vi sono inoltre due strati sinaptici dove i prolungamenti cellulari (assoni e dendriti), che per semplicità non sono stati rappresentati nell'illustrazione, si sovrappongono e stabiliscono reciproci contatti sinaptici.

La via primaria attraverso la retina è stabilita da tre tipi di neuroni: i fotorecettori, le cellule bipolari e le cellule gangliari, che si notano in questo ingrandimento. Una retina composta da queste cellule potrebbe trasmettere informazioni sulla luce, ma le cellule gangliari non esibirebbero proprietà selettive; darebbero solo una risposta analoga a quella da «centro» eccitato.

Le interazioni laterali possono spiegare l'esistenza di caratteristiche selettive, come la periferia antagonista di un campo recettivo. Si supponga che le cellule orizzontali producano sulle cellule bipolari un effetto (*frecce in nero*) opposto all'effetto dei fotorecettori. I segnali che alla fine sono trasmessi a una cellula gangliare avranno quindi un centro e una periferia antagonista. Le connessioni delle cellule orizzontali, che mediano questo effetto, non sono ancora del tutto note.

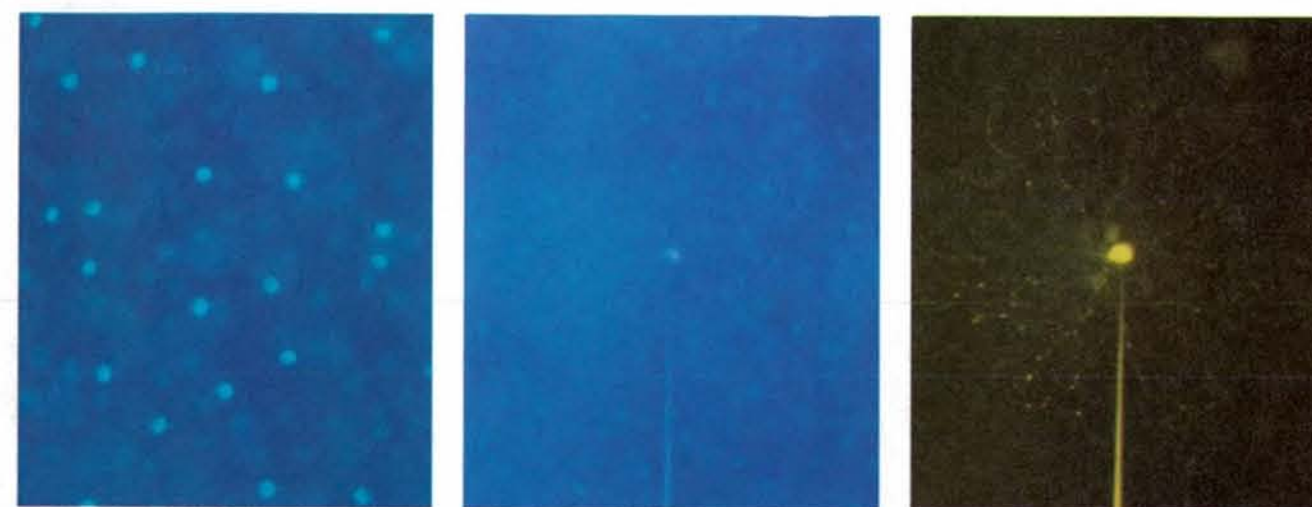
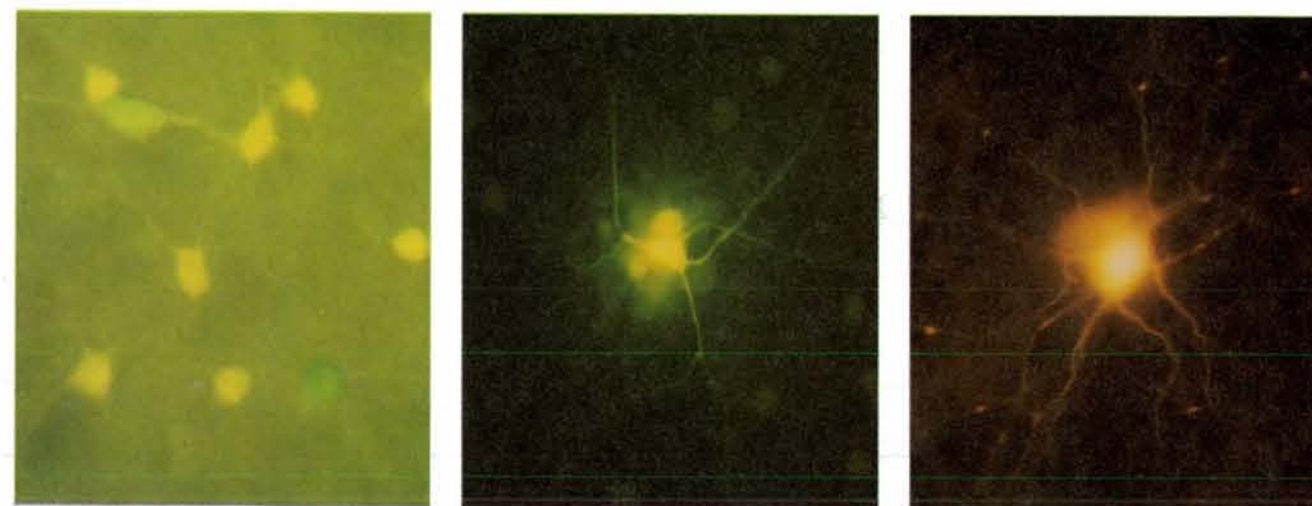
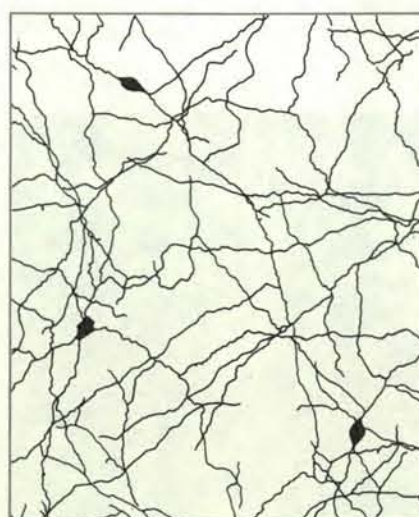
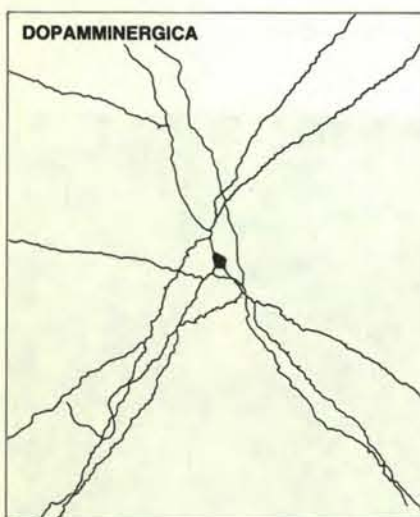
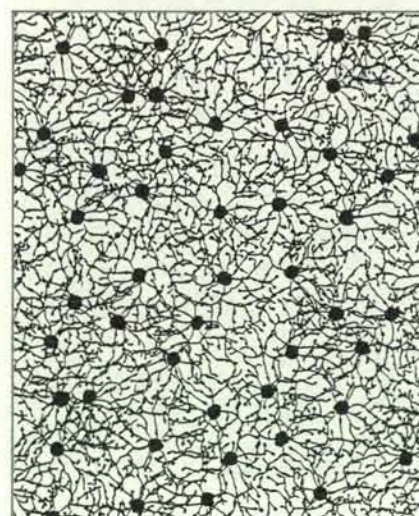
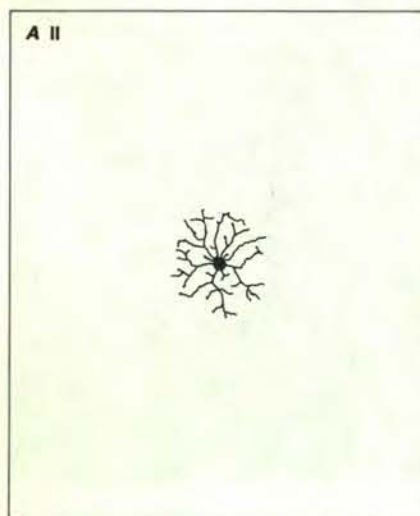
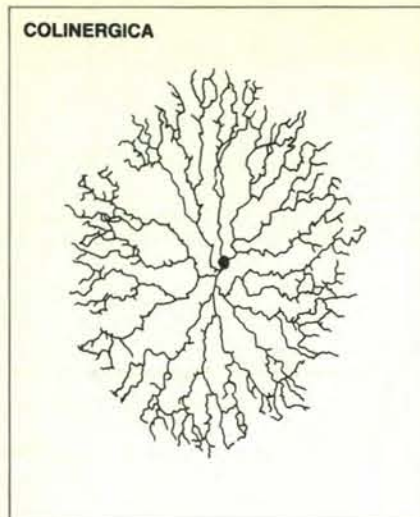
gio: l'acetilcolina che si trova nel tessuto viene immobilizzata in pochi millisecondi dal congelamento. Quando successivamente viene individuata mediante autoradiografia, deve trovarsi ancora all'interno della cellula alla quale apparteneva in origine. Vengono così eliminate in parte le incertezze che accompagnavano i primi metodi di identificazione dei neurotrasmettitori entro la cellula. La marcatura era netta e ben distinta: una cellula risultava o ben marcata o non marcata affatto. L'acetilcolina era presente solo in un piccolissimo sottoinsieme di cellule amacrine. Con il passar del tempo altri laboratori hanno confermato questa scoperta fondamentale di Ehinger. Come egli aveva suggerito, diversi neurotrasmettitori sono confinati in piccoli sottoinsiemi di cellule amacrine.

Conoscendo questo fatto, si è cercato di correlare neurotrasmettitori particolari a tipi morfologicamente distinti di cellule amacrine, come quelle descritte da Ramón y Cajal. Si è così trovato rapidamente che cellule amacrine con differenti forme dendritiche potevano essere appaiate a neurotrasmettitori particolari. Tra le varie prove a favore di questa ipotesi vi era anche il lavoro di Nicholas C. Brecha e Harvey J. Karten della State University of New York, a Stony Brook, dal quale risulta che sono presenti in cellule amacrine distinte non solamente i neurotrasmettitori tradizionali, ma anche molti peptidi di origine nervosa (cioè una vasta famiglia di neurotrasmettitori peptidici). Questa scoperta ha fatto aumentare oltre ogni aspettativa la diversità tra le cellule amacrine.

Il fatto che nelle cellule amacrine la forma e la biochimica della sinapsi fossero correlate tra loro ha avuto una seconda e importante conseguenza. Esso ha eliminato ogni dubbio residuo sull'esistenza di differenti funzioni biologiche in cellule amacrine dotate di forme diverse. Quando un neurotrasmettitore si fissa a un recettore su una cellula postsinaptica, questa cellula subisce mutamenti biochimici e fisiologici specifici. Se le varie cellule amacrine hanno forme differenti (il che implica diverse connessioni all'interno della retina) e differenti neurotrasmettitori, esse devono anche svolgere ruoli biologici distinti. Quale compito potrebbero avere 30 differenti cellule amacrine? Ciascun tipo indubbiamente ha una sua funzione specifica, ma sembrano esserci molti più tipi cellulari che compiti da eseguire. Nel mio e in altri laboratori alcune di queste funzioni incominciano ora a essere chiarite. Descriverò qui il lavoro compiuto con quattro cellule amacrine molto diverse.

La prima di queste cellule è stata già menzionata. Si tratta della cellula amacrina colinergica, cioè contenente acetilcolina, che è stata identificata mediante autoradiografia. Uno svantaggio dell'autoradiografia è però il fatto che la tecnica rivela il corpo cellulare di un neurone, ma non i suoi prolungamenti: assoni e dendriti che trasmettono e ricevono messaggi. Di fatto, il principale sistema usato in anatomia per rivelare i prolungamenti è la tecnica di Golgi, l'importante metodo usato da Ramón y Cajal nel XIX secolo. Esso non è però controllabile: colora un'intera cellula, ma il ricercatore non può sceglierla in anticipo.

Una tecnica sostitutiva di quella di Golgi è stata suggerita dalla mia scoperta che una molecola fluorescente, il 4,6--diammidino-2-fenilindolo (DAPI) si accumula selettivamente nei corpi cellulari dei neuroni colinergici. Se si tratta con

Un'ampia gamma di cellule amacrine presenti nella retina della carpa è rappresentata in questo disegno, pubblicato nel 1892 dal neuroanatomista spagnolo Santiago Ramón y Cajal, il quale riuscì a distinguere ben 14 tipi diversi di cellule esclusivamente in base alla loro forma.

Cellule amacrine colinergiche (il cui neurotrasmettitore è acetilcolina) sono state colorate con una tecnica messa a punto da Masaki Tauchi e dall'autore. Una retina intatta viene montata in piatto; i corpi delle cellule amacrine colinergiche accumulano in modo selettivo DAPI, un colorante fluorescente blu (*a sinistra*). Questo permette di fare arrivare una micropipetta riempita con un altro colorante, il giallo Lucifero CH, a una singola cellula amacrina illuminata in modo da rendere fluorescenti entrambi i coloranti (*al centro*). Con una differente illuminazione (*a destra*) solo il giallo Lucifero risulta fluorescente e rivela così il tipo di ramificazione che caratterizza i dendriti delle cellule colinergiche.

Altri due tipi di cellule amacrine colorati con il metodo descritto sopra. Una retina è stata trattata in modo che i corpi delle cellule dopamminergiche avessero fluorescenza verde, quelli delle cellule che accumulano indolammine fluorescenza gialla (*a sinistra*). L'iniezione del colorante in singole cellule mostra le specifiche conformazioni dendritiche delle cellule del primo (*al centro*) e del secondo tipo (*a destra*).

Mosaici di differenti cellule amacrine ricoprono diversamente la retina. Tre tipi di cellule sono presentati in una proiezione bidimensionale: a sinistra sono riportate cellule singole, mentre i disegni a destra rappresentano un'approssimazione del mosaico che risulterebbe se tutte le cellule di una classe si colorassero contemporaneamente. Le cellule colinergiche sono numerose e i loro dendriti ramificati formano un groviglio quasi ininterrotto (*in alto*). Il mosaico delle cellule amacrine AII è meno fitto (*al centro*), quello delle cellule dopamminergiche è ancora più rado (*in basso*). I disegni sono stati ricavati basandosi sulle caratteristiche medie delle cellule.

DAPI una retina *in toto* e poi la si monta in piatto, i neuroni colinergici emettono una fluorescenza marcata contro lo sfondo più scuro, il che ha permesso a Masaki Tauchi, che era rimasto a lavorare con noi dopo il dottorato, e a me di orientare una sottile micropipetta verso una cellula così da iniettare in questa un diverso colorante fluorescente, il giallo Lucifero CH. Il giallo Lucifero si diffonde attraverso l'intera rete di ramificazioni della cellula. I risultati sono stati perfettamente coerenti: tutte le cellule fluorescenti avevano la stessa forma e lo stesso tipo di ramificazione.

Il metodo non solo ci ha permesso di vedere la vera forma delle cellule colinergiche, ma anche di ricostruire il mosaico che queste cellule formano a mano a mano che coprono la retina. Per comprendere per quale motivo questo fatto è importante si deve di nuovo pensare al compito che la retina svolge. Innanzi tutto la retina deve salvaguardare la risoluzione spaziale; qualsiasi insuccesso comprometterebbe, infatti, l'acuità visiva. La densità e la disposizione spaziale degli elementi intermedi della retina, attraverso i quali il messaggio visivo deve passare, forniscono importanti informazioni sul ruolo che questi elementi possono svolgere all'interno del sistema nel suo insieme. Analisi pionieristiche di mosaici di cellule retiniche erano già state effettuate da Brian B. Boycott e Heinz Wässle al King's College di Londra per due cellule che possono essere colorate, come le popolazioni intere, con i metodi classici. I nostri metodi basati sulla fluorescenza permettono di studiare molte altre cellule.

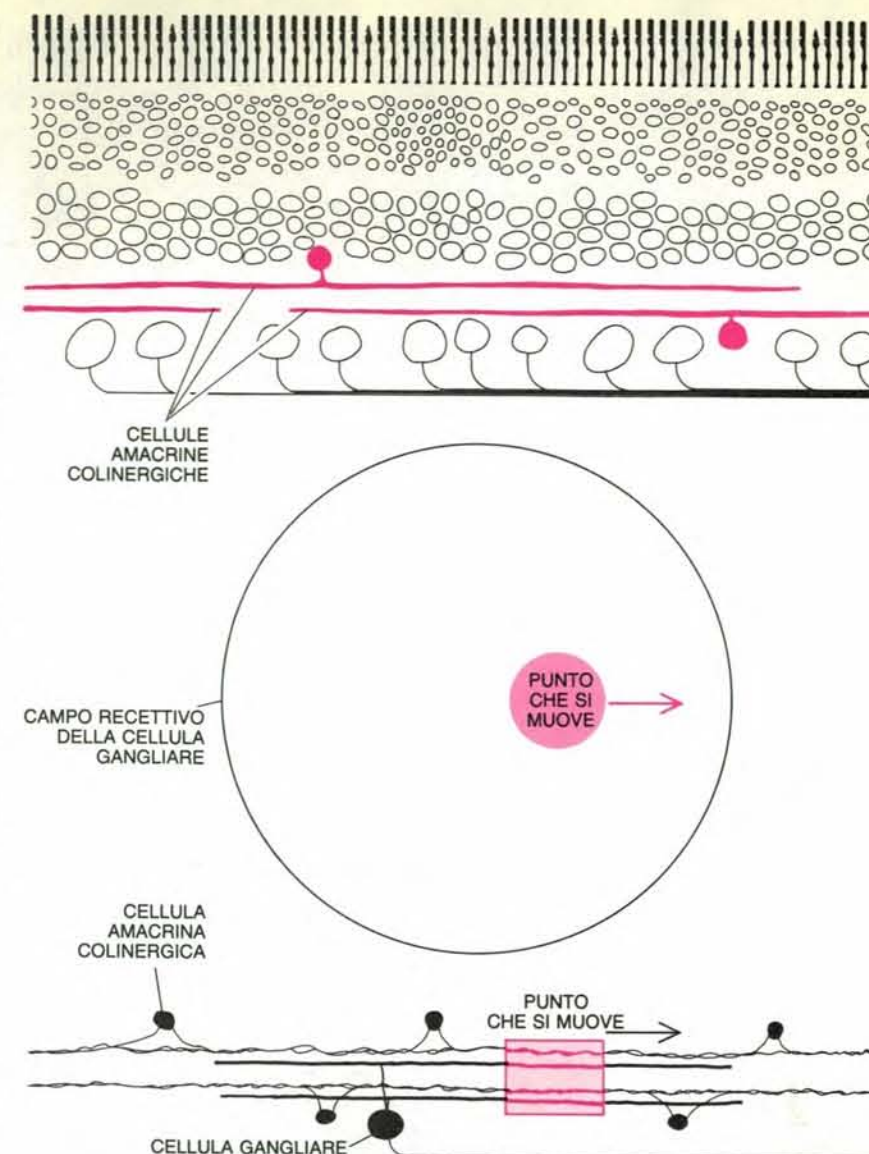
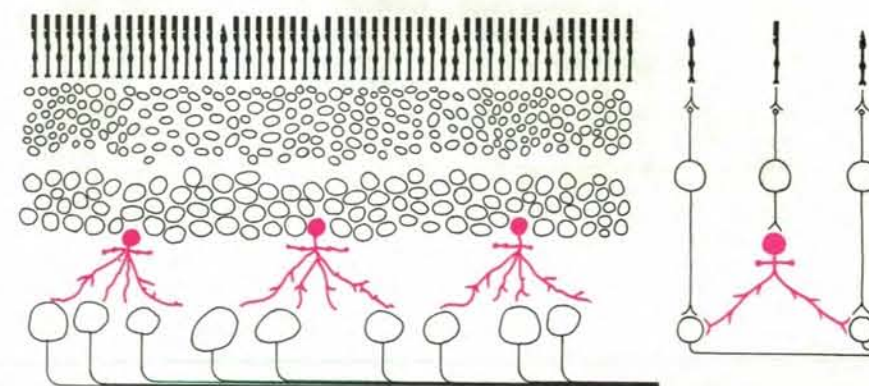
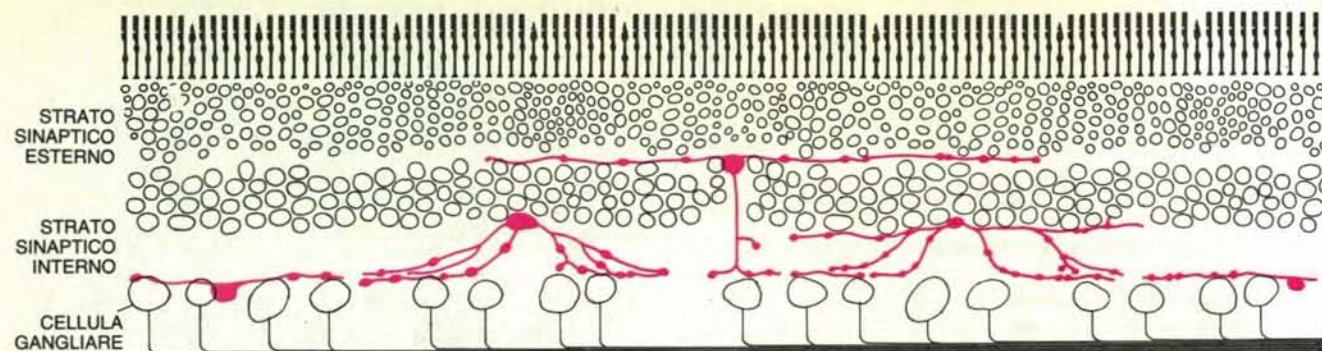
Quando Tauchi e io abbiamo misurato le dimensioni e la densità delle cellule colinergiche, abbiamo trovato un risultato sorprendente: le cellule si sovrappongono ampiamente. Nella porzione periferica della retina, dove la sovrapposizione è massima, un punto sulla superficie è rivestito da un intreccio di prolungamenti sovrapposti che provengono da qualcosa come 140 cellule amacrine colinergiche. Il lavoro di Boycott e Wässle aveva fatto pensare a una lieve sovrapposizione, ma una così enorme ridondanza era del tutto inattesa.

Una possibile spiegazione per tale ridondanza è stata suggerita dai tipi di stimoli che le cellule amacrine colinergiche possono distinguere. Da precedenti ricerche di elettrofisiologia sapevamo che le cellule amacrine colinergiche eccitano certe cellule gangliari, tra le quali vi sono quelle dotate di selettività direzionale. Sapevamo che queste cellule gangliari possono individuare il movimento di punti molto piccoli, addirittura inferiori all'area occupata dai loro dendriti. Tali punti sono anche molto più piccoli dell'area occupata dai dendriti delle cellule amacrine colinergiche. Siccome la maggior parte dei neuroni trasmette simultaneamente lo stesso messaggio attraverso tutte le sinapsi, ci si aspetterebbe

che una cellula amacrina liberi acetilcolina da tutte le sue ramificazioni dendritiche ogni qualvolta uno qualsiasi di questi punti viene eccitato. Come può una cellula così estesa trasmettere un'informazione precisa a proposito di stimoli tanto limitati rispetto a essa?

La migliore ipotesi che siamo riusciti a formulare è che i dendriti della cellula amacrina siano isolati elettricamente il che permetterebbe a una regione di scaricare localmente l'acetilcolina su una cellula gangliare, senza che alcuna scarica venga diretta verso punti più distanti. Questo fatto è coerente con la struttura delle cellule amacrine, nelle quali input e output sono presenti fianco a fianco sullo stesso prolungamento. L'attività elettrica delle cellule amacrine colinergiche consiste solo di potenziali elettrici graduati; non vengono generati invece impulsi nervosi, che si propagherebbero attraverso i prolungamenti dendritici. L'attività elettrica delle cellule amacrine colinergiche non dovrebbe quindi diffondersi molto al di là del punto di input. La prova diretta dell'esistenza di un simile meccanismo è di solito al di sopra delle nostre possibilità, ma sembra non esservi altro modo per spiegare la risoluzione di una cellula dotata di selettività direzionale. Il meccanismo rende inoltre molto vantaggioso il compatto intreccio di dendriti delle cellule amacrine colinergiche: un piccolo stimolo può essere percepito indipendentemente dal punto in cui perviene al campo recettivo della cellula gangliare.

Il secondo tipo di cellula amacrina che descriverò è la cellula AII, che differisce dalla cellula colinergica in quanto si estende lateralmente in misura molto limitata. David Vaney dell'Università di Cambridge, adottando il metodo basato sull'iniezione di sostanze fluorescenti, ha descritto il mosaico formato dalle cellule AII. Queste sono abbondanti e ricoprono l'intera superficie della retina, ma sono tanto piccole che i loro dendriti presentano poche sovrapposizioni.

Parecchi altri ricercatori hanno studiato le cellule AII: fra loro Helga Kolb e Ralph F. Nelson dei National Institutes of Health, Barbara A. McGuire, Peter Sterling e John K. Stevens della School of Medicine dell'Università della Pennsylvania e Ramon F. Dacheux e Elio Raviola della Harvard Medical School. Come risultato del loro diligente lavoro si può oggi dare una risposta a una domanda posta parecchi anni orsono dalla Kolb e da Edward V. Famiglietti: come è possibile che le cellule gangliari vengano eccitate da una fievole luce anche quando non ricevono un input sinaptico dalle cellule bipolari attivate dai fotorecettori a bastoncello? Sono solo questi i fotorecettori in grado di rispondere a una luce fievole. I gatti (come la maggior parte dei mammiferi) possono vedere anche in condizioni di scarsa illuminazione, eppure la loro anatomia prova che in essi le

Il potere risolutivo delle cellule amacrine colinergiche si può spiegare con il fatto che i dendriti sono prolungamenti sia per l'input sia per l'output e agiscono autonomamente rispetto alla cellula. Le cellule amacrine colinergiche appaiono in colore nella sezione della retina (*in alto*). I bordi di un punto che si sposta nel campo recettivo di una cellula gangliare (*al centro*) eccitano determinate regioni della ramificazione dendritica delle cellule amacrine (*in basso*); i dendriti delle cellule colinergiche eccitano pertanto la cellula gangliare solo nella zona illuminata.

Le cellule amacrine AII (*in colore*) sono situate tra le cellule bipolari stimolate dai bastoncelli e le cellule gangliari (*a sinistra*). Costituiscono così un ponte tra un tipo di cellula e l'altro (*a destra*). Nel disegno sono schematizzate le connessioni cellulari; in realtà, una cellula amacrina AII è collegata con diversi tipi di cellule gangliari attraverso vie nervose differenziate.

Le cellule che accumulano indolammine (*in colore*) possono presentare cinque morfologie differenti: quattro sono simili a quelle delle altre cellule amacrine; una è insolita in quanto costituisce un ponte tra gli strati sinaptici esterni e interni. Tutte le cellule, comunque, inviano i propri prolungamenti al livello inferiore dello strato sinaptico interno, nel quale la maggior parte di esse, o tutte, vanno a formare sinapsi con le cellule bipolari stimolate dai bastoncelli. Sembra che in questo modo esse forniscano cinque vie di accesso a un punto finale comune.

cellule gangliari ricevono uno scarso input o addirittura nessun input diretto da parte delle cellule bipolari attivate dai bastoncelli.

Un importante contributo al processo sopradescritto può provenire dalle cellule amacrine AII. Queste cellule hanno due funzioni ben precise: come molte altre cellule amacrine, trasmettono alle cellule gangliari una risposta «transitoria» alla luce, intensificando così la loro risposta iniziale alla stimolazione, ma servono anche a connettere le cellule bipolari, attivate dai bastoncelli, alle cellule gangliari e, facendo questo, permettono alle cellule gangliari di funzionare in condizioni sia di luce intensa, sia di luce fievole. Di fatto, le cellule amacrine AII divengono parte integrante della via che attraversa la retina. Il flusso dell'informazione procede dal bastoncello alla cellula bipolare, alla cellula amacrina AII e alla cellula gangliare; si capisce dunque perché le cellule amacrine AII siano piccole e stipate: queste caratteristiche consentono loro di trasmettere con la massima intensità lo stimolo lungo un percorso della massima importanza.

Le cellule del terzo tipo, come quelle del primo, sono contraddistinte dal loro neurotrasmettitore, in questo caso la dopammina. Tauchi e io abbiamo scoperto che potevamo usare su esse lo stesso tipo di sostanza impregnante fluorescente che avevamo usato sulle cellule colinergiche. Ciò che è notevole a proposito delle cellule dopamminergiche è la loro scarsa presenza nella retina. Per fare un esempio, una retina di coniglio è costituita in cifra tonda da qualcosa come 350 000 cellule gangliari, 300 000 cellule amacrine colinergiche e solamente 8500 cellule amacrine dopamminergiche. Inoltre i dendriti delle cellule dopamminargiche sono pochi e hanno ramificazioni sottili, il che dà luogo a un mosaico con molte lacune, non fittamente stipato come nel caso delle altre cellule amacrine di cui si è parlato.

Sebbene non si conosca la precisa funzione svolta dalle cellule dopamminergiche, il mosaico pieno di lacune fa pensare che queste cellule non siano interessate in attività che richiedono un elevato grado di risoluzione spaziale. Mentre il mosaico delle cellule amacrine colinergiche stipato in modo eccezionalmente compatto permette l'esatta risoluzione di un piccolo punto luminoso che attraversa la retina, nelle cellule dopamminergiche un analogo punto finirebbe spesso per cadere su uno dei vuoti del mosaico. In altre parole, è improbabile che le cellule dopamminergiche controllino una qualsiasi delle proprietà ovvie del campo recettivo delle cellule gangliari. Osservazioni al microscopio elettronico effettuate da John E. Dowling e Ehinger ad Harvard, permettono di rilevare che le cellule dopamminergiche hanno sinapsi solo con altre cellule amacrine e non hanno alcuna connessione diretta con la via primaria che attraversa la retina. Esperimenti di elettrofisiologia indicano, inoltre, che le cellule dopamminergiche influenzano l'attività delle cellule gangliari in modo vago, difficile da definire. Esse potrebbero avere, per esempio, una funzione piuttosto generica, come quella di regolare l'eccitabilità complessiva dei neuroni più interni della retina.

L'ultimo tipo di cellula amacrina che descriveremo in questo articolo si distingue sperimentalmente per il fatto che accumula analoghi chimici del neurotrasmettitore serotonina. Poiché la serotonina e i suoi analoghi sono indolammine, le cellule amacrine di questo tipo sono denominate «accumulatori di indolammine». Julie H. Sandell (che ha continuato a svolgere ricerca nel nostro laboratorio dopo aver conseguito il dottorato), e io stesso siamo stati in grado di esaminarle e di studiarne la morfologia dopo aver marcato la retina con un analogo fluorescente della serotonina, la 5,7-diidrossitriptammina.

Abbiamo potuto così stabilire che la suddetta classe di cellule amacrine può essere suddivisa in cinque tipi morfologici distinti, i quali però hanno caratteristiche in comune così numerose, che è meglio considerarli una famiglia di cellule piuttosto che tipi di cellule funzionalmente indipendenti. Vi sono molte ragioni per questo. In primo luogo, i dendriti hanno per dimensioni, forma e tipo di ramificazione una rassomiglianza familiare che li distingue dai dendriti delle altre cellule amacrine. In secondo luogo, e ancora più importante dal nostro punto di vista, partecipano tutti alla formazione di una fitta rete, o plesso, che tappezza il pavimento dello strato sinaptico più interno. In questa sede le cellule stabiliscono una sinapsi caratteristica, la cosiddetta sinapsi reciproca, con la terminazione di una cellula bipolare, attivata da un bastoncello.

Le differenze tra queste cellule si manifestano al di fuori del plesso e questo si riflette nelle loro connessioni sinaptiche e nella loro forma complessiva. Poiché le cellule che accumulano indolammine hanno estese ramificazioni e sono in contatto con molte, anche se non con tutte, le cellule bipolari attivate dai bastoncelli, si pensa che esse abbiano una grande influenza sulla via lungo la quale la radiazione luminosa di debole intensità attraversa la retina. La conclusione logica è che le cinque cellule in grado di accumulare indolammine rappresentano cinque differenti vie mediante le quali altri neuroni retinici possono interagire con i bastoncelli.

Come è messo in evidenza da questo piccolo campione di cellule amacrine, la retina contiene una notevole quantità di circuiti estremamente sofisticati. La complessità della retina è tale, di fatto, da sbigottire perfino i neurobiologi. «La retina viene ritenuta un sistema semplice» - si lamentano abitualmente i colleghi - «ma come potrebbe essere più complicata di così tra le varie strutture nervose?» Per stuzzicare il loro amor proprio si potrebbe rispondere che la retina sembra complicata solo perché di essa se ne sa di più (la paura inconfessata

Lo strato sinaptico interno è ingrandito circa 50 000 volte nella microfotografia elettronica (*in alto*) di R. F. Dacheux ed E. Raviola della Harvard Medical School. Si possono vedere i dendriti di un certo numero di neuroni, alcuni dei quali sono indicati sulla mappa (*in basso*). Il materiale scuro in un dendrite permette di identificare la cellula a cui quel dendrite appartiene come cellula che accumula indolammine, una tecnica messa a punto da J. H. Sandell e dall'autore. Essa ha permesso di trasformare il solito colorante fluorescente in un materiale opaco agli elettroni. In corrispondenza delle sinapsi si osservano vescicole (*cerchietti*) che contengono neurotrasmettitori (*frecce*).

è che la retina non sia in effetti particolarmente complicata e allora, se una struttura che viene ritenuta un prototipo di semplicità si presenta così intricata, che cosa sarà mai la corteccia cerebrale?). Si possono, comunque, trovare risposte più serie.

La prima è che il compito fondamentale della retina, la conversione di segnali luminosi in treni significativi di impulsi nervosi, richiede un meccanismo complesso. Questo articolo ha affrontato un solo aspetto della codificazione dell'input visivo da parte della retina: l'organizzazione di due tipi di campi recettivi delle cellule gangliari. Per esempio, ho completamente tralasciato di considerare la notevole capacità della retina di modificare i propri livelli di sensibilità, cosa che la retina fa mediante regolazioni interne della propria reattività, il che ha luogo a ogni stadio della gerarchia retinica. Sono queste le regolazioni che permettono alla maggior parte dei mammiferi di vedere bene sia alla luce del sole sia al chiarore delle stelle, un intervallo di intensità luminosa di 10 miliardi di volte. La retina è anche in grado di trasmettere al cervello informazioni riguardanti le varie lunghezza d'onda della luce che riceve, permettendo così agli esseri umani e a qualche altro vertebrato di vedere i colori. Date le funzioni sofisticate svolte dalla retina, si comprende perché le sue connessioni siano tanto complicate.

Una seconda risposta è che possono essere necessari circuiti sofisticati anche per quelle che sembrano funzioni semplici. Per esempio, non era possibile prevedere di trovare che i bastoncelli non formano sinapsi direttamente con le cellule gangliari e sembra illogico ritenere che il collegamento avvenga mediante le cellule amacrine AII. Non solo si tratta di una modalità indiretta di comunicazione, ma estende anche il percorso visivo (chiaramente di poco) in senso laterale, rischiando così di compromettere l'acuità visiva. Se i coni entrano direttamente in contatto con le cellule gangliari, perché dovrebbero comportarsi in modo differente i bastoncelli? L'unica conclusione che si può trarre è che la presenza di cellule amacrine AII può contribuire a risolvere alcuni problemi dei quali per ora non siamo nemmeno consapevoli.

La terza e forse la migliore risposta è che la retina riesce a condensare all'interno di una piccola area una grande complessità. Tutte le parti del sistema nervoso centrale operano con limiti di spazio, ma raramente lo spazio è così limitato come nella retina. Questo si verifica per due ragioni: le cellule retiniche devono essere stipate assieme in modo molto stretto allo scopo di rendere massima l'acuità visiva: inoltre la retina deve essere molto sottile (nel coniglio ha spessore di circa un decimo di millimetro) perché la luce possa penetrare fino ai coni e ai bastoncelli. Un millimetro quadrato della porzione centrale della retina di coniglio contiene ben 9,4 metri di soli dendriti di cellule colinergiche. Se a questi si aggiungono i dendriti di tutti gli altri neuroni presenti nella retina ci si rende conto della sbalorditiva densità dei prolungamenti cellulari. Non ci si deve quindi sorprendere che i neuroni retinici siano molto più piccoli della maggior parte degli altri neuroni: se i loro corpi cellulari avessero le stesse dimensioni di quelli contenuti dentro il cervello, l'occhio umano dovrebbe avere per lo meno le dimensioni di un mandarino. In poche parole, la retina rappresenta un autentico trionfo della miniaturizzazione più spinta.

La ricerca sulla struttura della retina sta per entrare in una nuova fase promettente: la ricostruzione in tre dimensioni di gruppi di neuroni collegati funzionalmente. I metodi che abbiamo messo a punto per rivelare la forma dei neuroni retinici si possono combinare facilmente con le tecniche della microscopia elettronica. Essi rendono più semplice il compito del microscopista fornendo una guida per seguire i singoli dendriti attraverso i loro addensamenti in ciascuno strato sinaptico. La stessa miniaturizzazione, che di solito rappresenta un ostacolo alla ricerca delle cellule retiniche, diviene un vantaggio quando ci si dedica alla ricostruzione tridimensionale di una struttura. Il fatto che le cellule retiniche siano densamente stipate permette di esaminare più facilmente interi circuiti retinici in microfotografie elettroniche di sezioni di tessuto in serie e di ricostruirne i collegamenti.

Sebbene questo tipo di ricerca sia per ora solo agli inizi, è evidente il valore dei modelli tridimensionali. Queste ricostruzioni del sistema nervoso rappresentano le descrizione più accurate della struttura e delle interconnessioni delle cellule retiniche. Ricerche di altro tipo saranno necessarie prima di poter dire come quelle stesse strutture funzionino sotto il profilo elettrico e biochimico. Quando finalmente saranno conosciute anche queste fondamentali caratteristiche, il lavoro iniziato quasi un secolo fa da Santiago Ramón y Cajal si potrà dire concluso.

Come le radici reagiscono alla forza di gravità

Anche se da oltre un secolo i botanici hanno scoperto che una radice disposta orizzontalmente si incurva verso il basso in risposta alla forza di gravità, solo ora il fenomeno comincia a essere meglio compreso

di Michael L. Evans, Randy Moore e Karl-Heinz Hasenstein

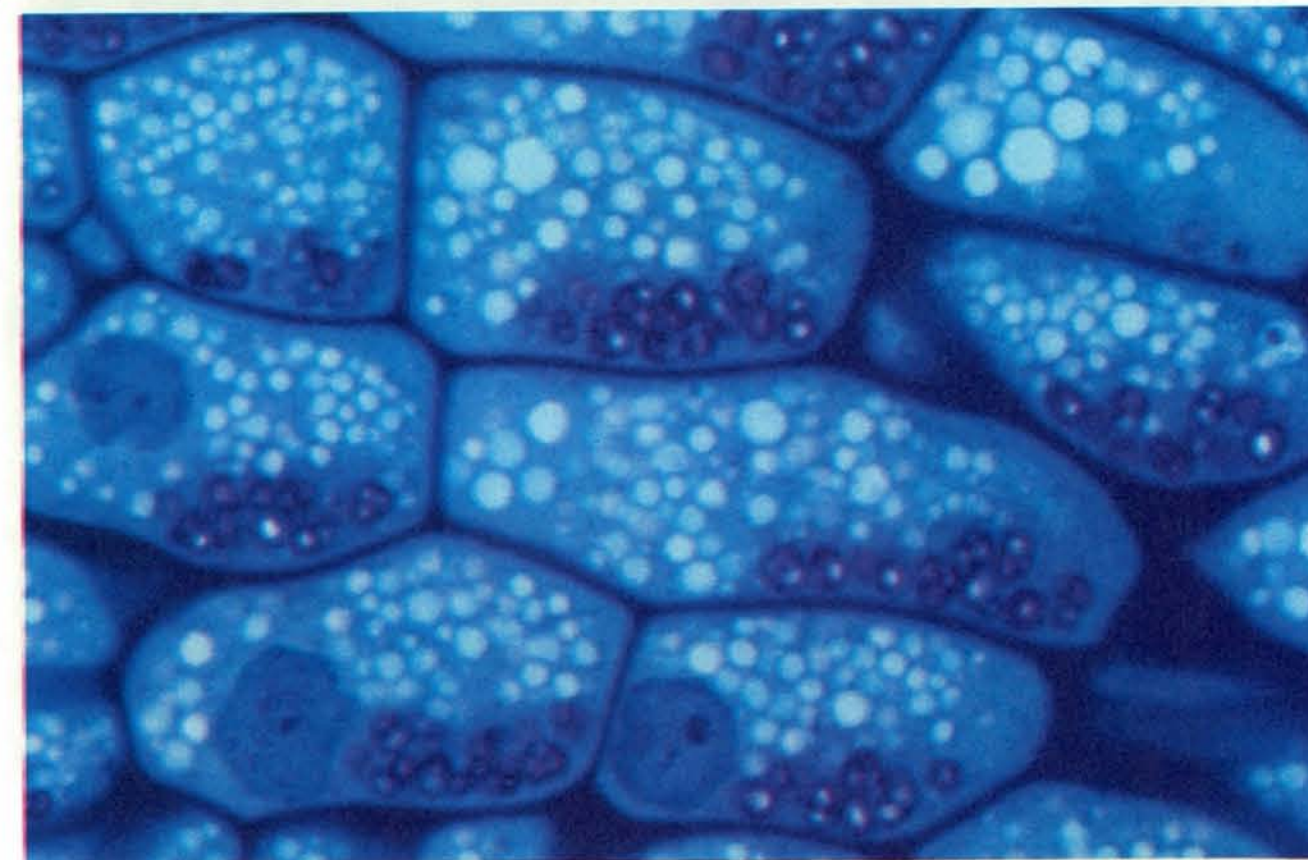
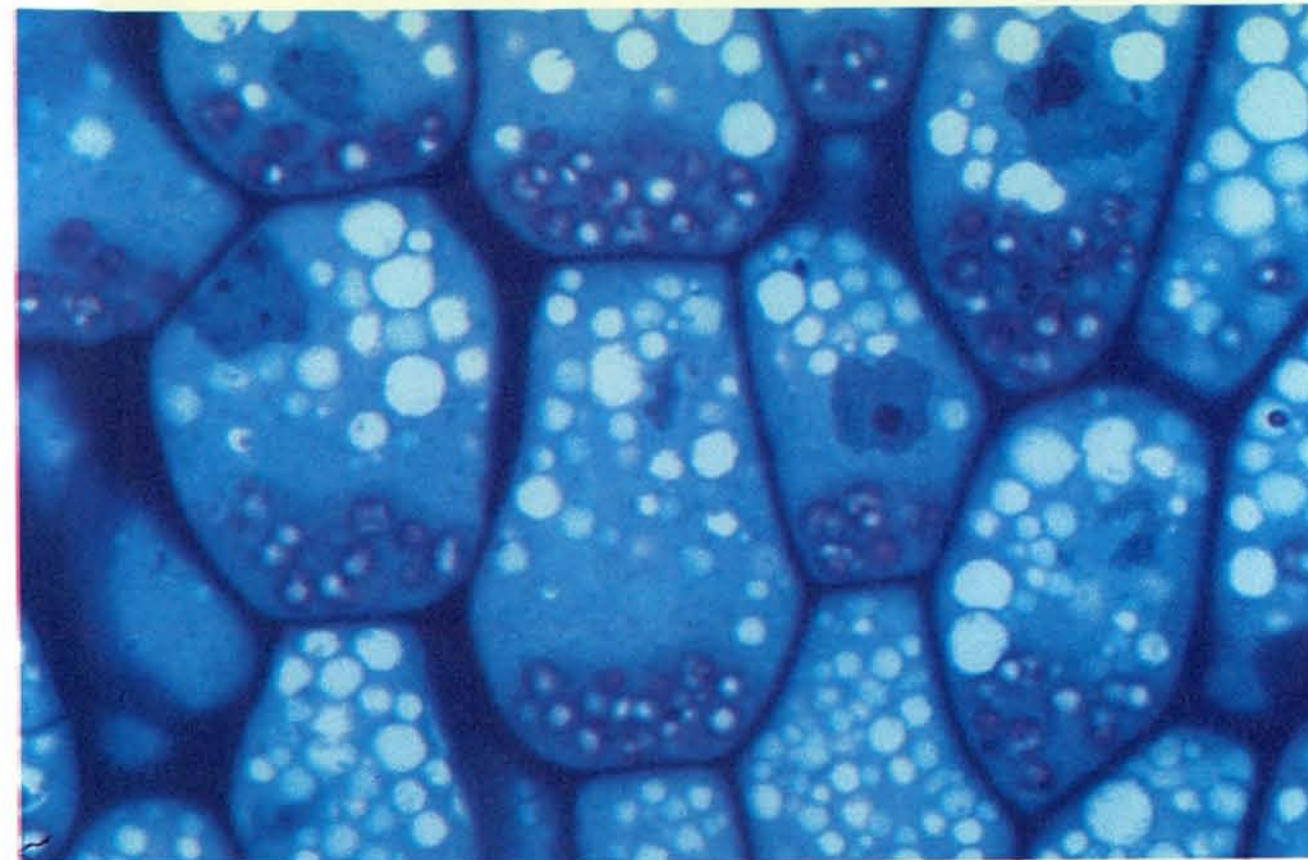
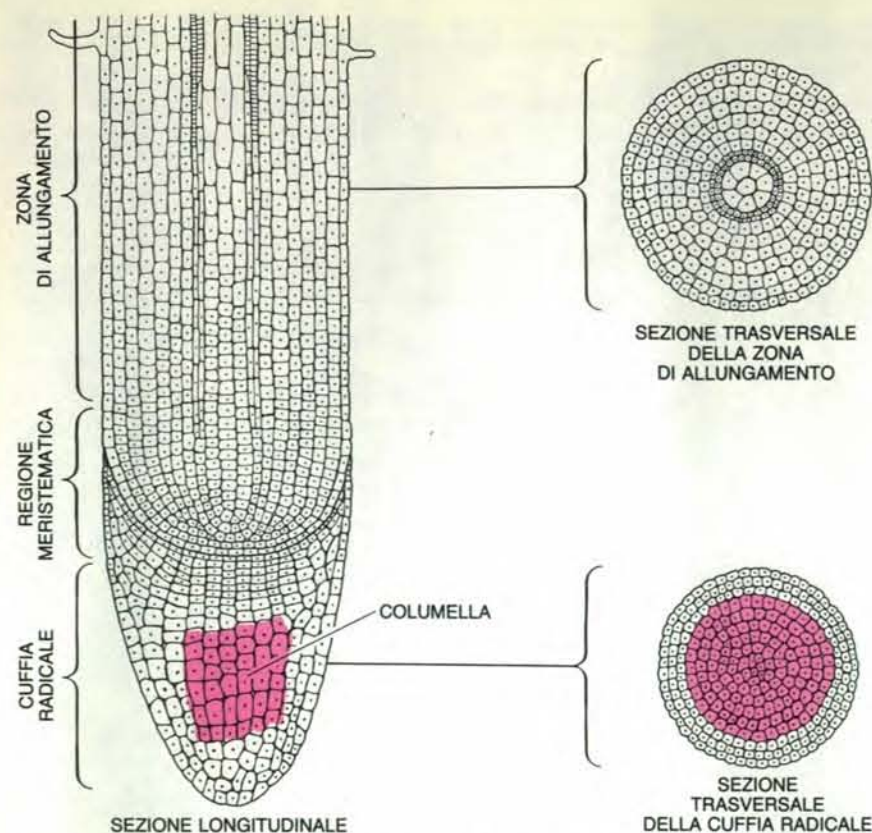
Nello spazio, dove non c'è gravità, non esiste neppure l'alto o il basso. Quando una pianta viene spostata da una certa posizione arbitrariamente considerata verticale a una posizione orizzontale, il fusto e la radice continuano a crescere diritti: non ritornano cioè alla loro posizione iniziale. Sulla Terra, invece, le piante mostrano un comportamento assai diverso noto come «gravitropismo» (fenomeno che una volta veniva chiamato «geotropismo»): le loro strutture si riorientano, per effetto d'una stimolazione gravitazionale (o «gravistimolazione»), disponendosi secondo il verso della forza di gravità. Se una pianta orientata verticalmente viene coricata su un fianco, l'apice del germoglio s'incurva verso l'alto, dalla parte opposta rispetto al verso della forza di gravità. Invece la radice, oggetto delle nostre ricerche, s'incurva verso il basso, ossia si orienta secondo la forza di gravità. In molti casi una radice dà inizio al processo dopo soli 10-30 minuti dalla stimolazione gravitazionale.

Un osservatore casuale potrebbe immaginare che questa curvatura delle radici verso il basso sia una semplice ricerca dell'acqua oppure una risposta passiva alla forza di gravità: ma la spiegazione è diversa. Le radici di alcune piante ricercano effettivamente l'acqua, ma la forza di gravità ha un effetto indipendente sul verso di crescita delle radici. Inoltre le radici in crescita, che hanno una forza sufficiente per penetrare in un suolo compatto, sono certamente in grado di resistere alla trazione esercitata dalla forza di gravità: il gravitropismo radicale deve avere pertanto un meccanismo diverso. I botanici avevano cercato di scoprire tale meccanismo già da prima del 1880, anno in cui fu pubblicato *The Power of Movement in Plants* (tradotto in italiano con il titolo *Il potere di movimento delle piante*) di Charles Darwin, che conteneva descrizioni particolareggiate del gravitropismo. Ciononostante, è solo con l'avvento del programma spaziale statunitense e con il relativo aumento degli investimenti nel campo della ricerca sugli effetti della forza di gravità e della gravità zero che i ricercatori hanno compiuto progressi significativi nella comprensione dei fenomeni del gravitropismo radicale.

Il gravitropismo si può dividere in tre fasi: percezione, trasduzione e risposta. Nelle radici la percezione dello stimolo gravitazionale sembra avvenire entro la cuffia, ossia nel mezzo millimetro che costituisce la parte terminale della radice. La risposta - un modello di crescita alterato che porta a una curvatura verso il basso - ha luogo a una certa distanza dalla cuffia, nella «zona d'allungamento», dove, come indica il termine, le cellule si allungano. La trasduzione, ossia la fase intermedia (che è la più misteriosa), quasi certamente implica un certo tipo di comunicazione tra la cuffia e la zona d'allungamento, ma solo ora si comincia a comprendere la natura esatta dell'interazione tra le due regioni.

Non è acora stato chiarito l'esatto meccanismo che consente alla radice di percepire la forza di gravità: ma già Darwin aveva riconosciuto che la cuffia radicale è il sito probabile di tale percezione. La cuffia, egli scriveva nel 1881, «avendo il potere di dirigere i movimenti delle parti attigue, agisce come il cervello di un animale inferiore: un cervello situato nell'estremità anteriore del corpo, che riceve impressioni dagli organi di senso e che dirige i diversi movimenti». Egli fu tra i primi a scoprire che l'asportazione della cuffia radicale elimina la capacità della radice di reagire alla forza di gravità. Dopo Darwin, altri ricercatori confermarono l'osservazione e dimostrarono che, sostituendo una cuffia asportata con una di un'altra radice, si ristabiliva il gravitropismo.

Secondo la maggior parte dei ricercatori, è la columella, ossia la regione situata in posizione centrale nella cuffia, a essere sensibile alla forza di gravità; questo tessuto è costituito da cellule contenenti amiloplasti, ossia organelli pieni di granuli d'amido. Nelle radici orientate verticalmente gli amiloplasti stanno nell'estremità inferiore di ciascuna cellula della columella, verso l'apice radicale. Nel giro di pochi secondi da quando le radici sono state stimolate gravitazionalmente (ossia coricate su un fianco), gli amiloplasti della columella si allontanano dalla posizione precedente e rapidamente si dispongono lungo la parete delle cellule che in quel momento si trovano in basso (*si veda l'illustrazione nella pagina a fronte*).

Si è osservato che lo spostamento degli amiloplasti costituisce lo stimolo iniziale che permette alla pianta di percepire il cambiamento di direzione rispetto alla gravità. Per esempio, le radici alle quali per esperimento siano stati asportati gli amiloplasti non reagiscono alla gravità. In alcuni casi l'inclinazione di una pianta su un lato per soli 12 secondi è uno stimolo sufficiente per dare inizio al riorientamento; gli amiloplasti appaiono i soli organelli della cuffia che risultino spostati in modo significativo nell'ambito di un periodo così breve. Nelle radici rese insensibili alla forza di gravità per asportazione della cuffia, la capacità di rispondere a questo stimolo è legata alla formazione e alla distribuzione entro la cellula di nuovi amiloplasti; le radici a cui sono state asportate le cuffie recuperano la sensibilità nei confronti della forza di gravità in un periodo compreso tra 14 e 22 ore, il tempo necessario perché gli amiloplasti si riformino e si ridistribuiscano in alcune cellule situate presso l'apice della radice amputata.

Si ritiene che le cellule della regione della columella, nella cuffia radicale, siano in grado di percepire la forza di gravità. Quando la radice è verticale (*in alto*), gli organelli opachi, detti amiloplasti (*corpi scuri*) si dispongono nella parte bassa delle cellule. Quando la radice viene disposta orizzontalmente (*in basso*), la forza di gravità obbliga gli amiloplasti a distribuirsi rapidamente sul nuovo lato inferiore delle cellule. Questo riassetto degli amiloplasti avvia una serie di eventi che fanno incurvare la radice verso il basso, nella direzione della forza di gravità.

Il gravitropismo si manifesta nelle parti terminali della radice. Le cellule in grado di percepire la forza di gravità si trovano nella regione centrale, detta columella, della cuffia radicale. La zona d'allungamento (ossia la zona di crescita), che spesso raggiunge una lunghezza di circa cinque-sei millimetri, è il sito dove si forma la curvatura gravitropica. La regione meristematica, situata tra le due zone precedenti, è quella in cui vengono prodotte le nuove cellule della radice.

Se questa ridistribuzione degli amiloplasti è effettivamente indispensabile per la percezione della forza di gravità da parte della radice, si tratta probabilmente dell'unica fase della risposta gravitropica in cui la forza di gravità attira direttamente verso il basso un determinato oggetto (in questo caso gli amiloplasti); le altre fasi sembrano essere mediate dall'attività chimica. Prima di esaminare queste successive fasi, dobbiamo segnalare che è stato descritto un mutante sviluppatosi in laboratorio che, pur non possedendo amiloplasti nella cuffia, presenta ancora una risposta gravitropica quasi normale. Ciò indica che il gravitropismo può avvenire anche in assenza di amiloplasti; inoltre, suggerisce che gli amiloplasti possono anche non essere le sole strutture predisposte per percepire la forza di gravità o che, in assenza di amiloplasti, le radici possono attivare un meccanismo sostitutivo di cui però non si conosce ancora nulla.

Qualunque sia il meccanismo implicato nella percezione gravitazionale, la reazione a questo stimolo, ossia la curvatura della radice, dipende chiaramente dalla crescita asimmetrica della zona d'allungamento. Quando una radice è orientata verticalmente, cresce uniformemente su tutti i lati. Se invece viene sistemata orizzontalmente, il lato superiore cresce più rapidamente di quello inferiore: questa crescita differenziata fa incurvare verso il basso la radice.

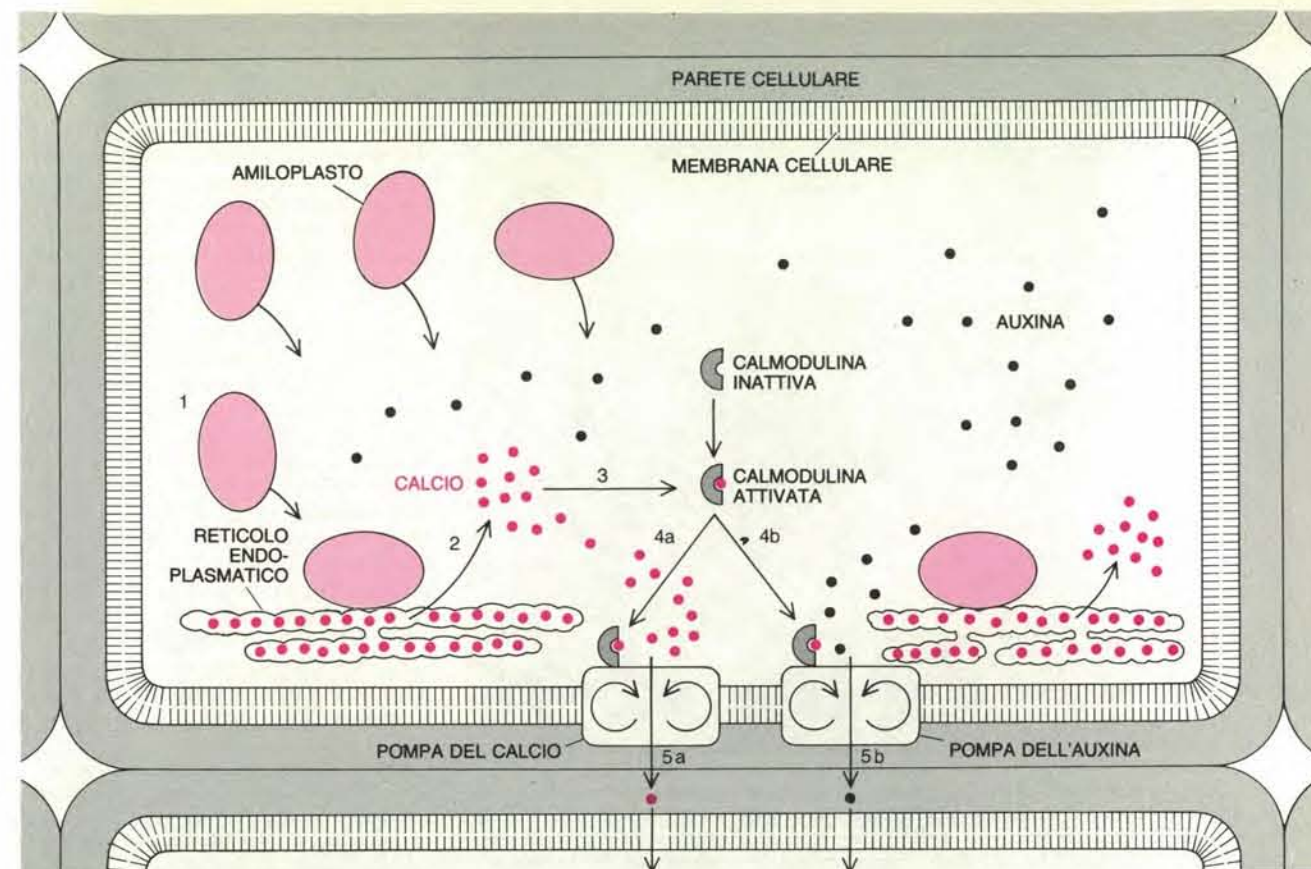
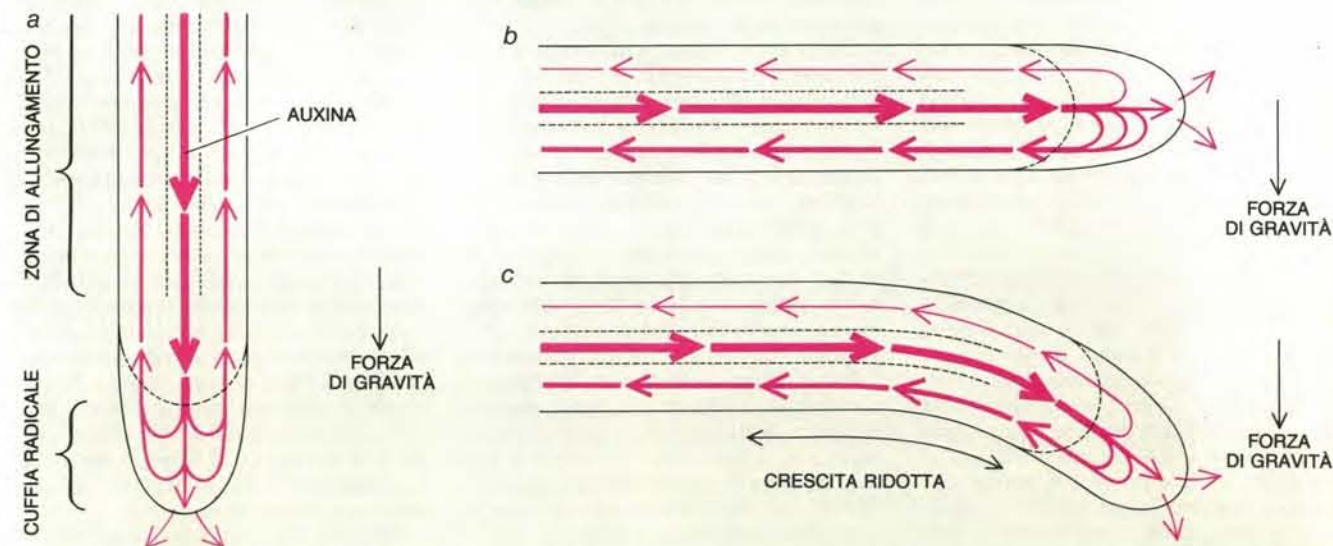
Per scoprire il fattore che altera direttamente il modello di crescita di una radice stimolata dalla forza di gravità, abbiamo provato, insieme ad altri ricercatori, a determinare i cambiamenti precisi di crescita lungo i lati superiore e inferiore di una radice disposta orizzontalmente. Abbiamo considerato svariate possibilità. Per esempio, ci siamo chiesti se l'aumento della velocità di crescita riguardi soltanto il lato superiore e non quello inferiore; se a diminuire sia la velocità sul lato inferiore mentre la velocità sul lato superiore aumenta; se la velocità di crescita aumenti su entrambi i lati della radice coricata, ma a un tasso più alto sul lato superiore, oppure, viceversa, se la velocità di crescita diminuisca su entrambi i lati, ma con un tasso più marcato sul lato inferiore.

Amy J. Nelson, nel corso di ricerche compiute in collaborazione con uno di noi (Evans), ha determinato il modello di crescita di radici di granoturco sottoposte a stimolazione gravitazionale. L'esperimento consisteva nel porre alcune pianticelle in una camera in grado di fornire l'umidità necessaria per la normale crescita della radice. In questo, come negli altri esperimenti presi in esame, ogni pianticella veniva sistemata in un ambiente che eliminava ogni possibilità di distribuzione irregolare di umidità, di luce, di sostanze nutritive o di temperatura con i relativi effetti sulla direzione di crescita della radice: grazie a questa precauzione si poteva essere sicuri che qualsiasi curvatura nella radice sarebbe stata attribuibile alla sola forza di gravità. La Nelson sistemava la radice bianca di ciascuna pianticella davanti a uno sfondo bianco e applicava granelli neri lungo la superficie della radice, che è spontaneamente appiccicosa. Una telecamera collegata a un calcolatore veniva poi focalizzata sulla radice così marcata. Il calcolatore seguiva il movimento graduale dei granelli a mano a mano che la radice si sviluppava valutando la velocità di crescita di varie parti della radice prima e durante la risposta gravitazionale.

La Nelson ha riscontrato che la stimolazione gravitazionale provoca una riduzione nella velocità di crescita delle superfici della radice, ma che la riduzione più marcata si verifica lungo la superficie inferiore, in una zona situata da due a tre millimetri di distanza dalla cuffia. Altri ricercatori, con differenti metodi sperimentali e con altri tipi di piante, hanno trovato analogamente che la crescita ritardata sul lato inferiore della radice crea una curvatura verso il basso nelle radici stimolate dalla forza di gravità.

Quale potrebbe essere la ragione del rallentamento della crescita lungo la superficie inferiore? La crescita delle cellule vegetali è controllata in larga misura da ormoni: ciò fa pensare a un aumento di concentrazione di qualche ormone inibitore della crescita lungo la superficie inferiore della zona di allungamento. Secondo un'ipotesi sul gravitropismo radicale, l'inibitore in questione potrebbe essere l'acido abscisico, un ormone che si forma naturalmente nelle radici delle piante. Quando l'acido abscisico viene applicato in forti dosi sulla superficie esterna delle radici, viene assorbito e ne inibisce la crescita. Nelle radici di certe pianticelle di granoturco che non mostrano gravitropismo al buio, l'illuminazione della cuffia radicale induce gravitropismo entro un'ora; durante tale periodo si è trovato che aumenta anche la velocità di sintesi dell'acido abscisico nella cuffia radicale.

Nonostante questi dati sperimentali, le conclusioni che abbiamo potuto trarre sulla base degli esperimenti più recenti ci hanno indotto a respingere l'ipotesi dell'acido abscisico. Uno di noi (Evans) ha dimostrato infatti che l'ormone inibisce l'allungamento della radice solo quando viene applicato alle radici a concentrazioni significativamente superiori rispetto a quelle che si pensa si verifichino in natura.

Un dato ancor più significativo è stato

Secondo il meccanismo proposto, la distribuzione degli amiloplasti nella cuffia radicale sottoposta a stimolazione gravitazionale è collegata allo spostamento verso il basso degli ioni calcio (*pallini in colore*) e dell'auxina (*pallini in nero*): tale spostamento dipende dalla calmodulina, un importante enzima che attiva altri enzimi cellulari. Quando si mette una radice su un fianco, gli amiloplasti delle cellule della columella si ammassano sul lato inferiore di ogni cellula, venendo a contatto con il reticolo endoplasmatico, un complesso di membrane ricche di calcio (*1*). La pressione sul reticolo induce il complesso di membrane a liberare calcio nel citoplasma (*2*). Quando il livello di calcio nel citoplasma lungo il lato inferiore della cellula raggiunge una certa soglia, il calcio si lega alla calmodulina attivandola (*3*): a sua volta questa sostanza attiva due tipi di enzimi presenti nella membrana cellulare del lato inferiore della cellula: una pompa del calcio (*4a*) e una pompa dell'auxina (*4b*). Tali pompe trasportano rispettivamente il calcio e l'auxina in eccesso entro la parete cellulare (*5a e 5b*). Successivamente il calcio e l'auxina, per mezzo di meccanismi diversi, migrano attraverso le cellule sottostanti verso il lato inferiore della cuffia radicale.

Lo schema di spostamento dell'auxina (*frecce*) in una radice cambia quando questa è sottoposta a stimolazione gravitazionale. In una radice verticale (*a*), l'auxina migra, attraverso la parte centrale, dalla zona di allungamento alla cuffia. Dopo che l'auxina è entrata nella cuffia, si ritiene che una parte di essa si sposti lungo la cuffia ed esca attraverso l'apice; nel frattempo un'altra parte di auxina si distribuirebbe sui lati e poi rifluirebbe simmetricamente verso la zona di allungamento. Se si orienta orizzontalmente la radice (*b*), il modello del flusso di auxina diviene asimmetrico. Gli autori propongono che la maggior parte dell'auxina che penetra nella cuffia si sposti verso il lato inferiore, dove il calcio accumulatosi fa aumentare la velocità con cui l'auxina ritorna verso la zona di allungamento. L'eccesso di auxina che si costituisce nelle cellule situate nella metà inferiore della zona di allungamento ne inibisce la crescita portando a un incurvamento della radice (*c*).

L'importanza della cuffia per il gravitropismo radicale è dimostrata dalla differenza di crescita tra due radici sottoposte a stimolazione gravitazionale, delle quali una è priva di cuffia (*in alto*) e l'altra ha la cuffia intatta (*in basso*). Se la cuffia viene asportata, la radice cresce normalmente, ma non s'incurva più verso il basso quando viene sistemata orizzontalmente. Per contro, una radice che presenti la cuffia intatta, dopo essere stata riorientata, si incurva verso il basso.

rilevato da un altro di noi (Moore): le radici delle pianticelle di granoturco cresciute in presenza di un inibitore della sintesi di acido abscisico, pur non contenendo concentrazioni di acido rilevabili, si incurvano verso il basso in risposta a una stimolazione gravitazionale. Analogamente, le radici di un mutante del granoturco incapace di sintetizzare l'acido abscisico sono in grado di reagire alla forza di gravità. Ancora Evans ha potuto osservare che le radici immerse in una soluzione a elevata concentrazione di acido abscisico s'incurvano verso il basso, dopo essere state sottoposte a stimolazione gravitazionale. Anche questo è un dato significativo, perché è improbabile che un basso gradiente di concentrazione ormonale indotto dalla forza di gravità possa stimolare da solo una curvatura, quando la radice viene immersa in soluzioni sature della sostanza.

Appaiono molto più convincenti le prove a favore dell'ormone auxina quale fattore in grado di controllare la curvatura radicale gravitropica. L'auxina, che come l'acido abscisico si trova naturalmente nelle radici, è un potente inibitore della loro crescita, anche a concentrazioni da 100 a 1000 volte inferiori rispetto a quelle in cui l'acido abscisico mostra la sua efficacia. Nelle radici rese insensibili alla forza di gravità (per asportazione della cuffia radicale), l'auxina, applicata in piccola dose su un lato della zona d'allungamento, può indurre le radici a incurvarsi verso il lato dove essa è stata applicata: grazie a questo semplice dato sperimentale, è ragionevole pensare che un piccolo aumento fisiologico di auxina sul lato inferiore di una radice disposta orizzontalmente basti per far incurvare la radice stessa verso il basso.

Dopo aver applicato auxina marcata con isotopi radioattivi in modo uniforme sulla zona d'allungamento di una radice sottoposta a stimolazione gravitazionale, si è osservato lo spostamento dell'ormone verso il lato inferiore della radice: l'esperimento fa pensare che almeno un poco dell'auxina naturale si possa spostare per raccogliersi sulla parte inferiore della zona di allungamento. Inoltre, si è osservato che le sostanze chimiche che nelle radici interferiscono con lo spostamento dell'auxina da una cellula all'altra e che presumibilmente impediscono l'aumento di deposizione di auxina nell'area della curvatura gravitropica, inibiscono la capacità della radice di incurvarsi per effetto della forza di gravità. Per dimostrare in modo più decisivo il ruolo svolto dall'auxina entro la radice sarebbe necessario verificare se, in natura, questo ormone diventa più concentrato nella parte inferiore della zona di allungamento, appena prima dell'inizio della curvatura gravitropica. Ma l'esperimento non è ancora stato eseguito.

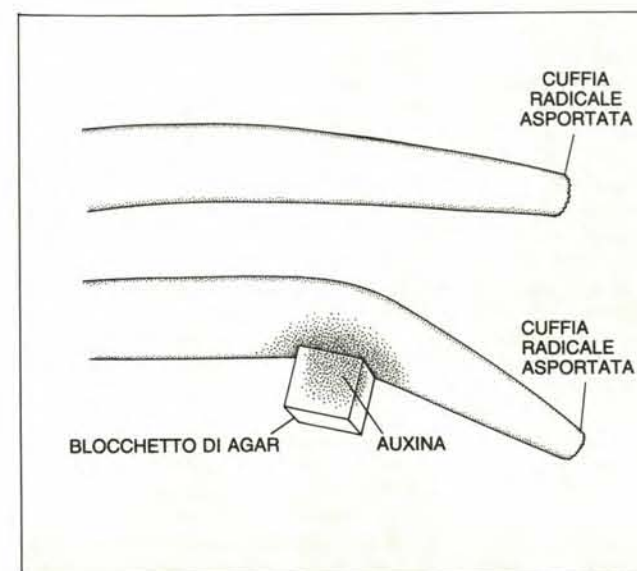
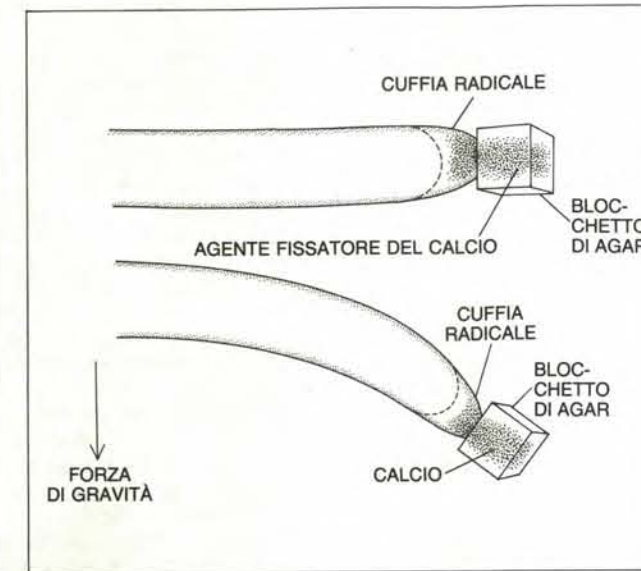
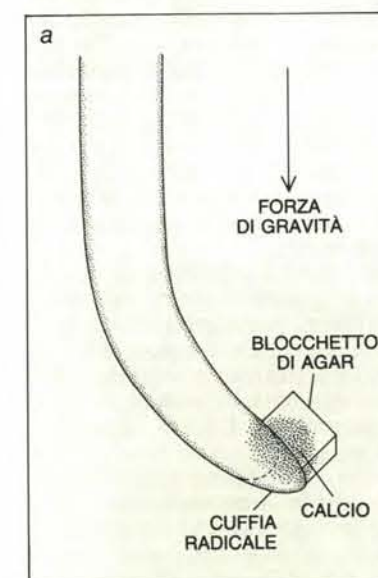
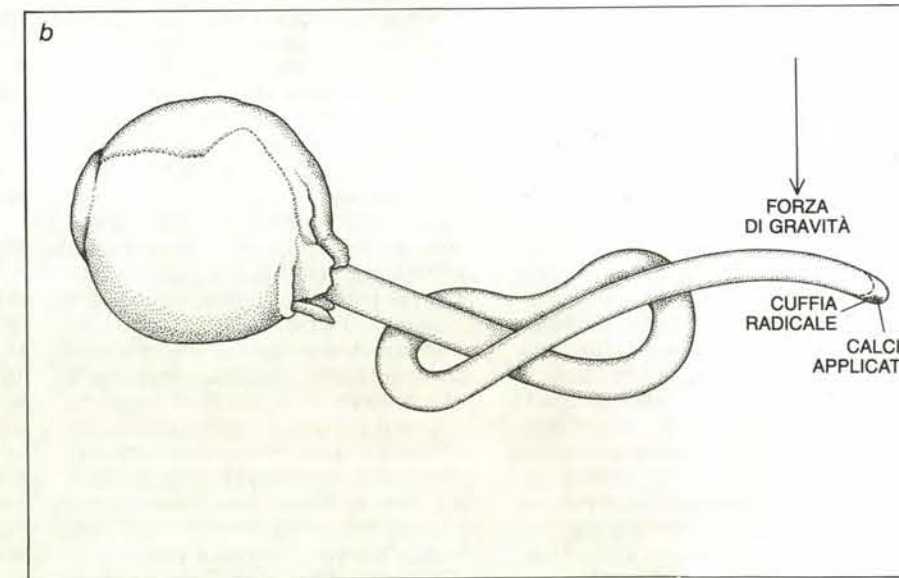
Se la distribuzione degli amiloplasti nella columella della cuffia radicale influenza la percezione della forza di gravità e se un aumento di auxina sul lato inferiore della zona d'allungamento induce in modo definitivo la reazione di curvatura della radice verso il basso, qual è il nesso che unisce questi due fenomeni? In altre parole, in che cosa consiste la misteriosa fase intermedia di trasduzione del gravitropismo?

Sulla base dei dati recentemente raccolti, la spiegazione da noi proposta è la seguente: la distribuzione degli amiloplasti scatena la liberazione di ioni calcio (Ca^{++}) da parte di organelli situati lungo il lato inferiore delle cellule della columella; a sua volta, il calcio che viene così liberato attiva i sistemi di trasporto che trasferiscono il calcio e l'auxina verso il basso da una cellula all'altra, in direzione del lato inferiore della cuffia radicale. Quando il calcio è nella parte inferiore della cuffia, può facilitare il movimento di auxina dalla cuffia verso il lato inferiore della zona di allungamento.

Il nostro gruppo è stato il primo a formulare l'ipotesi che la liberazione e lo spostamento del calcio immagazzinato potessero essere gli importanti passaggi iniziali della fase di trasduzione del gravitropismo radicale, dopo che parecchi articoli pubblicati negli anni sessanta e settanta avevano riportato che, quando i germogli reagiscono alla forza di gravità, il calcio appare spostarsi da una cellula all'altra. In esperimenti condotti su germogli sottoposti a stimolazione gravitazionale, si era potuto dimostrare che il calcio migrava verso la superficie superiore del fusticino appena prima che il germoglio s'incurvasse verso l'alto.

Per verificare la nostra ipotesi, Konrad M. Kuzmanoff e Timothy J. Mulkey, in collaborazione con Evans, applicarono l'acido etilendiamminotetraacetico (EDTA), una sostanza chimica che si lega al calcio e lo immobilizza, alle cuffie delle radici di granoturco. Con nostra sorpresa, le radici trattate con EDTA, pur continuando a crescere a velocità normale, divennero totalmente insensibili alla forza di gravità. Asportando l'EDTA e sostituendolo con calcio, le radici ridivennero sensibili allo stimolo gravitazionale. Con questo esperimento, quindi, il gruppo di ricerca riuscì a dimostrare che il calcio libero e mobile nella cuffia radicale è essenziale per il gravitropismo della radice. L'osservazione è stata ulteriormente confermata da altre ricerche che hanno dimostrato che, dopo alcuni trattamenti compiuti allo scopo di impedire il movimento del calcio, il gravitropismo può essere del tutto annullato.

Riprendendo in esame gli esperimenti che Kuzmanoff e Mulkey avevano interrotto, June S. Lee, dell'Università di Stato Kang Reung, in Corea, durante il periodo trascorso presso l'Ohio State University è riuscita a stabilire la direzione probabile dello spostamento del calcio all'interno delle radici sottoposte a stimolazione gravitazionale. June S. Lee ha dapprima studiato radici di controllo che non erano state stimolate scoprendo che il calcio ha un'influenza notevole sulla loro direzione di accrescimento. Quindi, applicando in modo disuniforme il calcio alla cuffia radicale, ha notato che la radice s'incurvava verso l'area dove il calcio era più concentrato. Anzi, dopo un'applicazione ininterrotta di calcio su un lato della cuffia di una radice orientata verticalmente, la radice finiva per incurvarsi facendo un giro completo di 360 gradi. Sapendo che le radici stimolate dalla gravità s'incurvano verso il basso, la Lee ha concluso che tale curvatura poteva derivare da un eccesso di calcio probabilmente depositato sulla superficie inferiore della cuffia per uno spostamento di ioni calcio verso il basso.

In seguito la Lee ha dimostrato che il calcio poteva effettivamente compiere questo tipo di movimento all'interno della cuffia; ella applicava uniformemente calcio radioattivo sulle radici di granoturco, orientava le radici sia verticalmente, sia orizzontalmente e quindi

Se si dispongono orizzontalmente radici di granoturco a cui sia stata asportata la cuffia si dimostra che la crescita asimmetrica nella zona di allungamento potrebbe essere provocata da una distribuzione sbilanciata dell'auxina. L'asportazione della cuffia rende le radici insensibili alla forza di gravità (*in alto*); se però si applica un blocchetto di agar riempito di auxina sul lato di una radice (*in basso*), la radice cresce più lentamente lungo quel lato e si incurva dove è stato applicato l'ormone.

Una radice di granoturco disposta orizzontalmente, sulla cui cuffia sia stato applicato un fissatore del calcio, continua a crescere orizzontalmente (*in alto*): la sostanza, infatti, impedisce la ridistribuzione del calcio in risposta alla forza di gravità. Sostituendo al blocchetto di agar con l'agente fissatore un blocchetto con calcio, il gravitropismo radicale si ricostituisce (*in basso*). Ciò dimostra che la mobilità del calcio nella cuffia è fondamentale per fare reagire la radice alla forza di gravità.

Le radici di granoturco a cui sia stato applicato calcio si incurvano verso il lato dell'applicazione. Appena dopo aver sistemato un blocchetto di agar contenente calcio sul lato destro della cuffia di una radice verticale (*a*) si è osservato un incurvamento verso destra. Analogamente, variando la localizzazione del calcio applicato (*b*), si sono ottenute parecchie curvature e ciascuna rappresenta un inarcamento della radice verso il calcio. Tali esperimenti indicano che la distribuzione irregolare del calcio, o più precisamente l'eccesso di calcio lungo il lato inferiore della cuffia sottoposta a stimolazione gravitazionale, contribuisce all'incurvamento della radice verso il basso nella zona di allungamento.

misurava lo spostamento degli ioni radioattivi. Nelle radici orientate verticalmente il calcio radioattivo rimaneva distribuito uniformemente, mentre nelle radici coricate su un fianco il calcio radioattivo risultava spostato verso il lato inferiore della radice. Questo movimento direzionale del calcio era particolarmente rilevante nella cuffia radicale.

Abbiamo trovato nuove conferme all'ipotesi dello spostamento verso il basso del calcio nelle radici stimolate gravitazionalmente in un tipo di ricerca del tutto diverso. Disponendo speciali microelettrodi accanto a una radice di crescione orientata verticalmente, H. M. Behrens e collaboratori dell'Università di Bonn hanno osservato che lungo la superficie della radice e all'interno della radice stessa, presso l'apice, si aveva un flusso simmetrico di corrente elettrica. Quando questi ricercatori sistemarono la radice in posizione orizzontale, trovarono che l'andamento della corrente diveniva asimmetrico: in particolare, il flusso della corrente lungo la superficie superiore della cuffia radicale era diretto verso l'esterno, attraverso l'apice, mentre lungo la superficie inferiore era diretto verso l'interno. Gli esperimenti indicavano che la corrente veniva trasportata da un flusso di ioni idrogeno (H^+).

Thomas Björkman e A. Carl Leopold della Cornell University hanno recentemente confermato queste osservazioni studiando radici di granoturco. Essi hanno trovato che due-sei minuti dopo che le radici erano state sistemate orizzontalmente, si verificava un drastico cambiamento nella direzione della corrente nella cuffia radicale: il flusso della corrente era diretto verso l'alto e verso l'esterno attraverso l'apice della cuffia. Björkman e Leopold hanno formulato l'ipotesi che il flusso verso l'alto di ioni idrogeno potrebbe indirettamente corrispondere a un flusso di ioni calcio verso la parte inferiore della cuffia; per mantenere la neutralità elettrica, la cuffia dovrebbe bilanciare questo spostamento di ioni calcio con un controflusso di altre cariche positive, come gli ioni idrogeno. (L'intensità della corrente diretta verso l'alto misurata da Björkman e Leopold è in accordo con le nostre misure a proposito degli spostamenti di ioni calcio verso il fondo della cuffia radicale.) Mentre la corrente verso l'alto corrisponde alla migrazione di ioni calcio, la collocazione temporale del fenomeno fa pensare che venga provocato dal riassetto degli amiloplasti: infatti il trasferimento del calcio avviene immediatamente dopo questo processo e precede la curvatura della radice verso il basso.

Le modalità attraverso le quali il riassetto degli amiloplasti nella cellula scatena il movimento degli ioni calcio che vanno poi ad accumularsi lungo il bordo inferiore della cuffia radicale non sono ancora state chiarite. Tuttavia Barbara G. Pickard della Washington University a St. Louis ha proposto un'interessante ipotesi che è in accordo con i risultati di Björkman e Leopold. Nelle cellule della columella, gli amiloplasti spostati dalla stimolazione gravitazionale si depositano sopra il reticolo endoplasmatico, un complesso di membrane ricche di calcio e di vescicole. La Pickard suggerisce che la pressione degli amiloplasti sul reticolo endoplasmatico faccia sfuggire dal complesso gli ioni calcio provocando un aumento della concentrazione di tale elemento sul bordo inferiore delle cellule.

Basandoci su questo modello, proponiamo un'ipotesi più avanzata. Ci sembra verosimile che, quando il livello di calcio raggiunge una determinata soglia, gli ioni attivino la calmodulina, una piccola proteina nota come potente attivatore di molti enzimi importanti per il funzionamento della cellula, non solo nelle piante, ma anche negli animali e persino in alcuni microrganismi. Appena attivata dal calcio, la calmodulina sembra stimolare lo spostamento del calcio mettendo in funzione le relative «pompe», enzimi formati da grandi molecole disposte attraverso la membrana cellulare. Le pompe attivate sul bordo inferiore delle cellule espellono allora il calcio in eccesso che finisce con l'accumularsi lungo il bordo inferiore della cuffia radicale.

La calmodulina postulata da questo modello sembra effettivamente essere presente nella cuffia radicale. Charles L. Stinemetz, uno studente che lavorava con uno di noi (Evans), ha dimostrato che tale sostanza è presente nelle radici di granoturco e che la concentrazione di calmodulina nella cuffia è quadrupla rispetto alla zona d'allungamento. A riprova del ruolo potenziale della calmodulina, Stinemetz scoprì anche che tale sostanza nella cuffia può essere necessaria per il gravitropismo radicale. Quando applicò inibitori della calmodulina alle cuffie delle radici, osservò un forte ritardo della curvatura gravitropica.

Allo scopo di spiegare le fasi successive della risposta gravitropica - lo spostamento verso il basso dell'auxina nella cuffia radicale e lo spostamento, stimolato dal calcio, dell'auxina verso il bordo inferiore della zona di allungamento - dobbiamo prima definire qual è il modello più probabile di trasferimento dell'auxina attraverso la radice. Nelle radici orientate verticalmente, l'auxina si sposta verso la cuffia radicale attraverso la parte centrale della zona di allungamento. Una parte dell'auxina in movimento nella cuffia viene metabolizzata oppure si trasferisce, attraverso l'apice, nel mezzo di crescita. Nello stesso tempo un po' di auxina si distribuisce verso i lati della cuffia, dove viene riinviata verso l'alto, all'interno della zona di allungamento, a opera delle cellule situate nella parte esterna della radice *(si veda l'illustrazione in basso a pagina 65)*. Lo spostamento è simmetrico, per cui la radice verticale cresce diritta.

Nelle radici sottoposte a stimolazione gravitazionale si osserva una variazione nello schema di spostamento dell'auxina nella cuffia radicale. Sulla base dei dati recentemente raccolti da uno di noi (Hasenstein), possiamo affermare che la calmodulina, attivata dal calcio presente nelle cellule della columella, mette in funzione non solo le pompe del calcio nella parte inferiore della membrana cellulare, ma anche le pompe dell'auxina (di cui è nota la presenza nelle membrane delle cellule vegetali). Riteniamo che queste pompe facciano uscire l'auxina dalle cellule della columella, attraverso il lato inferiore, cosicché molta auxina che entra nella cuffia radicale proveniendo dalla zona di allungamento migra verso il lato inferiore della cuffia, anziché dividersi simmetricamente tra i lati superiore e inferiore. È nostra opinione, inoltre, che l'elevata concentrazione di calcio all'interno delle cellule lungo il lato inferiore della cuffia radicale faccia aumentare in qualche modo la velocità con cui l'auxina, in quella parte della cuffia radicale, ritorna nel lato inferiore della zona di allungamento.

Riprendiamo ora il quesito posto più di un secolo fa: qual è il meccanismo per cui una radice sistemata orizzontalmente reagisce alla forza di gravità? La nostra idea parte dalla constatazione che la forza di gravità attira gli amiloplasti nelle cellule della columella della cuffia radicale. Gli amiloplasti, ammassandosi sul reticolo endoplasmatico del lato inferiore delle cellule, fanno aumentare i livelli di calcio nel citoplasma; ciò attiva la calmodulina la quale, a sua volta, fa entrare in azione le pompe del calcio e dell'auxina nella membrana situata sul lato inferiore delle cellule. Le pompe espellono il calcio e l'auxina attraverso il lato inferiore della cuffia radicale. La concentrazione elevata di calcio in questa parte della cuffia fa quindi aumentare l'auxina lungo la via che porta questo ormone verso il lato inferiore della zona di allungamento della radice. In tale sito l'auxina inibisce notevolmente la crescita e quindi induce la radice a incurvarsi verso il basso nella stessa direzione della forza di gravità.

È ovvio che il modello della crescita della radice verso il basso, indotta dal gravitropismo porta quasi sempre la radice verso le riserve d'acqua e di sali nutritivi. Le nostre recenti scoperte e quelle degli altri botanici confermano l'ipotesi, avanzata da Darwin, secondo la quale questo comportamento autoconservativo della radice sarebbe controllato dalla cuffia. Attualmente, gli studiosi impegnati nelle ricerche sul gravitropismo radicale e sugli altri modelli di comportamento della radice dispongono di nuove prove che non possono che riecheggiare l'affermazione di Darwin del 1881: «Per quanto concerne le funzioni, non c'è alcuna struttura delle piante più meravigliosa dell'apice della radice.»

Le eruzioni vulcaniche esplosive

La quasi totalità delle 30 000 vittime dovute ad attività vulcanica negli ultimi 10 anni ha perso la vita a causa di fenomeni connessi con eruzioni esplosive come quella del Mount St. Helens del 1980 o di intensità inferiore

di Roberto Scandone

Il 20 marzo 1980, un terremoto di magnitudo 4 con ipocentro localizzato sotto il vulcano St. Helens, negli Stati Uniti, segnò l'inizio di una serie di eventi che sarebbero culminati il 18 maggio 1980 con una catastrofica eruzione esplosiva. Il numero di lavori sul Mount St. Helens pubblicati dopo il 1980 è impressionante e il vulcano è certamente divenuto uno dei meglio conosciuti al mondo. Infatti, non appena ebbe inizio l'attività del vulcano, l'US Geological Survey installò una schiera di sensori in grado di registrare anche le minime variazioni di attività di un vulcano che si risvegliava dopo 123 anni di quiescenza. L'analisi dei dati raccolti ha fornito un quadro di quella che è la struttura del vulcano e dell'evoluzione della sua attività eruttiva.

L'eruzione è stata al tempo stesso la verifica e il punto di partenza per lo sviluppo delle nuove teorie sui meccanismi eruttivi di cui parleremo nell'articolo.

La classificazione delle eruzioni esplosive

Una prima basilare suddivisione delle eruzioni è quella fra eruzioni effusive (cioè quelle che emettono prevalentemente colate di lava come avviene per esempio nell'Etna o nel Kilauea) ed eruzioni esplosive durante le quali il magma viene frammentato prima di giungere in superficie e viene emesso sotto forma di pomici, lapilli, bombe e ceneri (prodotti piroclastici). Questo secondo tipo di eruzioni, molto più violente delle prime, è caratterizzato da una più efficiente trasformazione dell'energia termica del magma in energia meccanica.

Il grado di esplosività di un vulcano è da porre in relazione al differente contenuto in silice dei magmi e al contenuto di acqua juvenile disciolta in essi: un alto contenuto in silice rende il magma più viscoso e impedisce che l'acqua ne sfugga durante il processo di risalita.

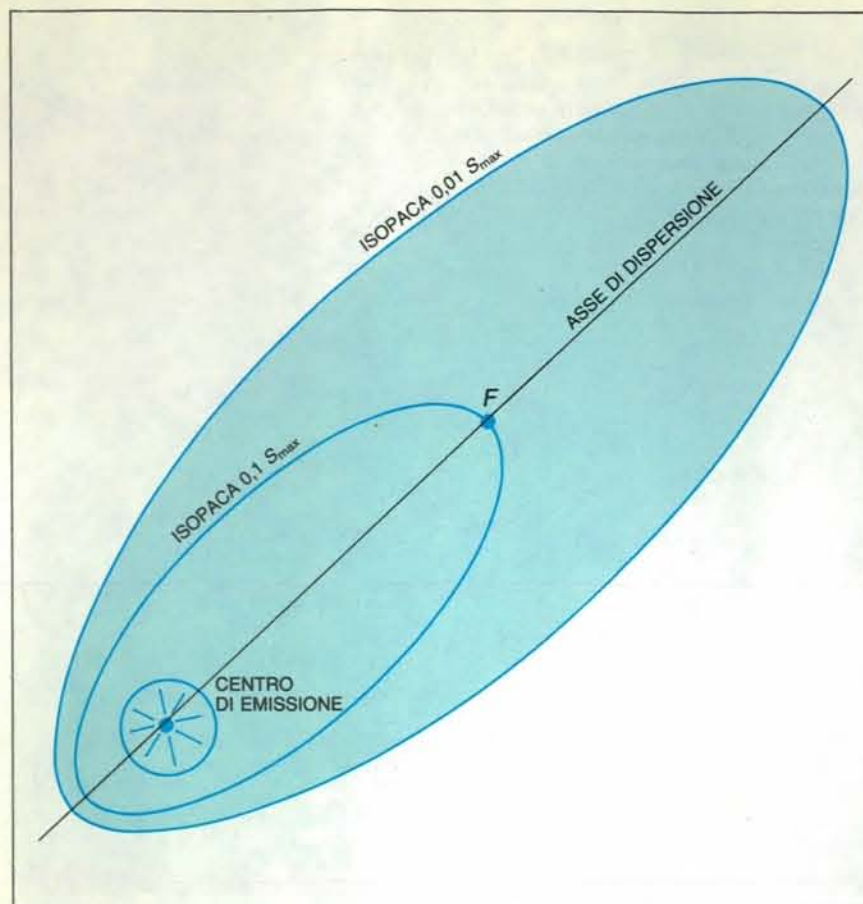
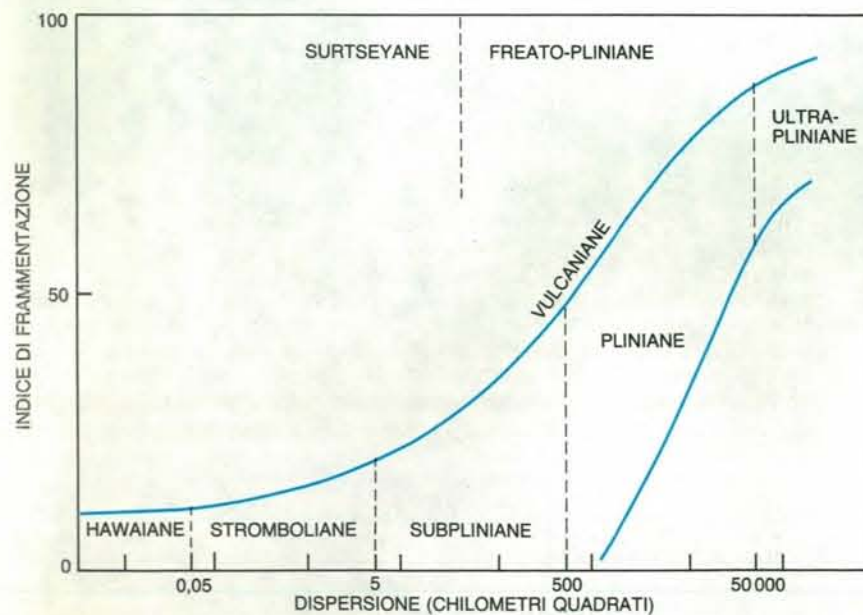
Uno schema di classificazione delle eruzioni basato unicamente sul volume dei prodotti emessi non dà un quadro completo della «violenza» di un'eruzione. Questa è la ragione per cui la classificazione delle eruzioni, e in particolare di quelle esplosive, viene fatta utilizzando sia il volume dei prodotti emessi sia altri parametri che descrivono le modalità con cui esse avvengono.

Intorno alla metà degli anni sessanta, l'inglese George P. L. Walker, che attualmente lavora presso l'Hawaii Institute of Geophysics, ha proposto uno schema particolare di classificazione delle eruzioni esplosive basato su due indici: l'indice di frammentazione (F) e l'indice di dispersione (D).

Durante un'eruzione esplosiva i prodotti piroclastici lanciati in aria e trasportati dal vento ricadono a terra formando depositi di spessore decrescente con la distanza dal centro di emissione. Il luogo dei punti di uguale spessore è definito «isopaca»; le isopache hanno generalmente una forma ellissoidale con il vulcano che occupa uno dei fuochi. L'indice di dispersione D è l'area delimitata dall'isopaca con valore uguale a 1/100 dello spessore massimo misurato in prossimità del vulcano. L'indice di frammentazione F è la percentuale (in peso) dei prodotti che sono più fini di un millimetro. Esso si misura in un campione raccolto nel punto in cui l'asse di dispersione dei prodotti interseca l'isopaca di 1/10 dello spessore massimo.

Nello schema di classificazione proposto da Walker (*si veda l'illustrazione in basso a pagina 72*) i vari tipi di eruzione occupano campi che corrispondono a differenti valori di F e di D, con una variazione progressiva di F e di D nel passaggio da eruzioni hawaiane (F e D minimi) a ultrapliniane (F e D massimi). Le eruzioni hawaiane sono caratterizzate da una bassa esplosività, con scarsa efficienza di conversione dell'energia termica in energia meccanica. Le eruzioni ultrapliniane hanno un alto grado di esplosività e di efficienza di conversione dell'energia termica. Le eruzioni surtseyane, freato-pliniane e vulcaniane si discostano dall'aumento progressivo di D e di F. Infatti il valore di F è più alto di quello che corrisponderebbe al normale campo di variazione. Questa caratteristica è dovuta al differente meccanismo eruttivo che si ha quando il magma interagisce con acqua di origine esterna. La maggior esplosività che ne deriva provoca una maggiore frammentazione e un conseguente elevato valore di F.

In termini pratici questa classificazione fornisce una misura del grado di esplosività dell'eruzione e una indicazione del meccanismo con il quale essa avviene. D'altra parte questo schema di classificazione si basa sulla distribuzione di quelli che vengono definiti «prodotti di caduta» (*air-fall tephra*). In molte eruzioni esplosive, invece, una percentuale rilevante di prodotti si deposita con meccanismi definiti di «flusso piroclastico». In questo caso una miscela di gas e di prodotti piroclastici, avente una densità superiore a quella dell'atmosfera circostante, scivola lungo i fianchi del vulcano come una vera e propria valanga. È necessario pertanto introdurre un nuovo parametro, chiamato «violenza» dell'eruzione, che consente di valutare la velocità alla quale si propagano i flussi piroclastici. La misura della «violenza» è data dal rapporto fra lo spessore medio del deposito di un flusso piroclastico e il raggio del cerchio che ha area uguale a quella ricoperta dal deposito. Quanto più è elevata la velocità di propagazione del flusso piroclastico tanto più è basso questo rapporto. In altre parole, si ha un deposito molto sottile esteso su una

Le fotografie del Mount St. Helens, prima e dopo l'eruzione del 18 maggio 1980, documentano l'ampiezza del nuovo cratere. La fotografia in alto è stata ripresa da Robert Decker nel giugno 1970, mentre quella in basso, di Ray Foster dei Sandia Laboratories, è del luglio 1980. Il Mount St. Helens era inattivo dal 1857. I fenomeni precursori del parossismo del 18 maggio sono durati per due mesi. L'eruzione ha rimosso l'intero fianco settentrionale del vulcano con la formazione di un cratere a forma di ferro di cavallo con un diametro di circa due chilometri.

I prodotti di caduta di un'eruzione esplosiva si distribuiscono generalmente su un'area di forma ellissoidale (*in colore chiaro*) il cui asse, chiamato «asse di dispersione», coincide con la direzione opposta a quella da cui spirano i venti dominanti. Le linee di uguale spessore dei prodotti di caduta sono chiamate «isopache». L'area delimitata dall'isopaca che racchiude i prodotti il cui spessore è maggiore di 0,01 volte lo spessore massimo (S_{max}) è l'indice di dispersione (D). L'indice di frammentazione, indicato con F, è la percentuale in peso dei prodotti che sono più fini di 1 millimetro misurata nel punto di intersezione fra l'isopaca 0,1 S_{max} e l'asse di dispersione.

La classificazione proposta da George P. L. Walker per le eruzioni esplosive si basa sugli indici di dispersione e di frammentazione. Le eruzioni che cadono nell'area compresa tra le due curve sono eruzioni magmatiche; nelle altre si ha un grado diverso di interazione con acqua esterna. La scala orizzontale è logaritmica perché la dispersione varia di molti ordini di grandezza.

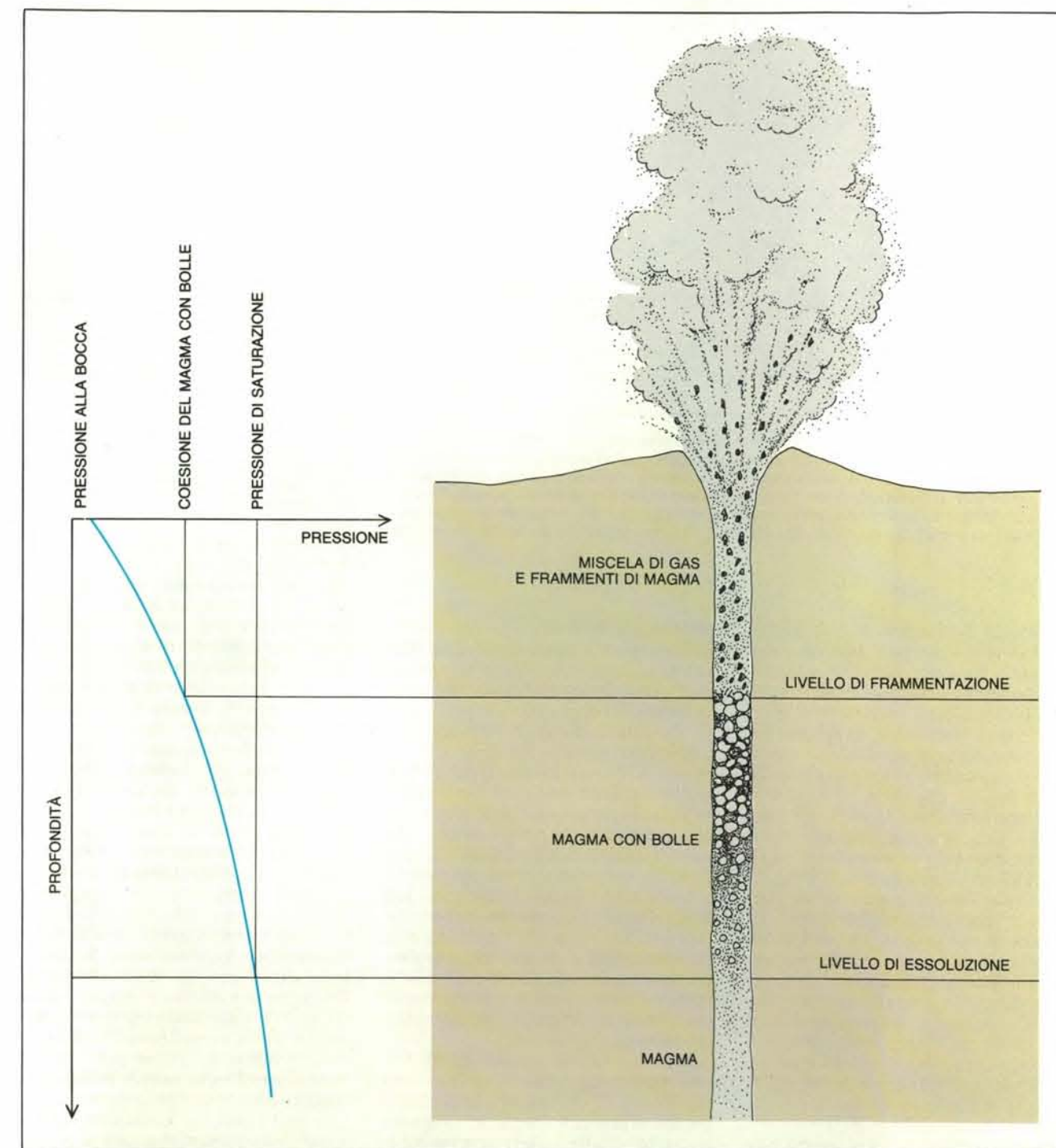
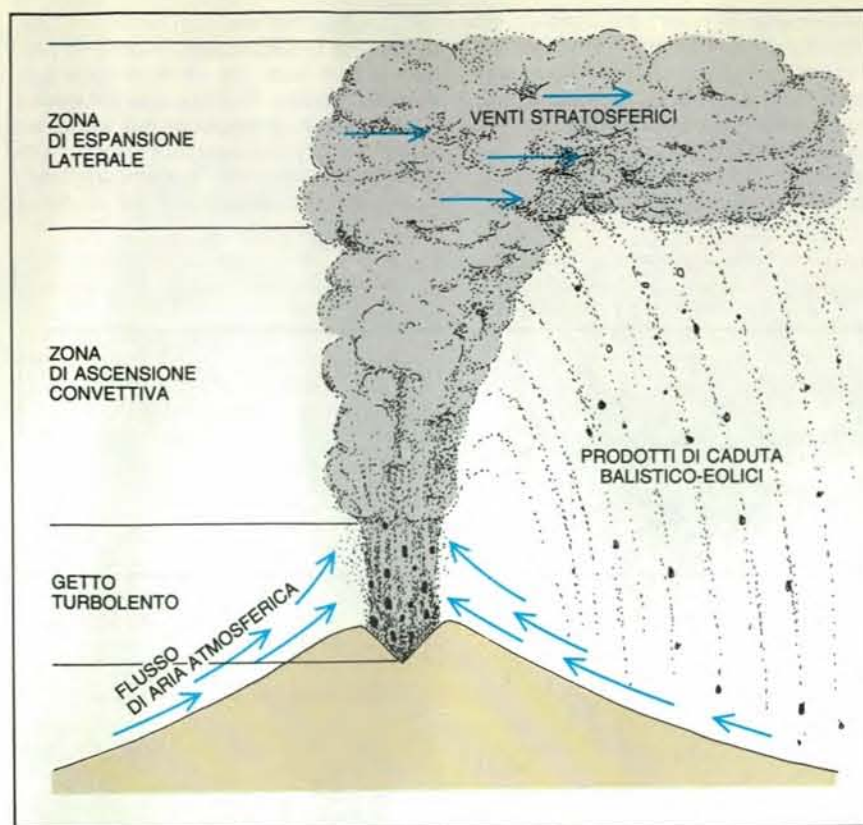
grande superficie. I valori tipici per eruzioni molto violente sono compresi fra 1/1000 e 1/70 000. L'ultimo valore è stato misurato per i prodotti di un'eruzione molto violenta avvenuta 1800 anni fa in Nuova Zelanda presso il lago di Taupo.

Lo studio teorico dei meccanismi eruttivi

Una teoria fisica complessiva che tenga conto dei vari fattori che entrano nella dinamica di una eruzione è ben lungi dall'essere elaborata; tuttavia importanti passi si vanno compiendo in questa direzione. Gli inglesi Lionel Wilson, dell'Università di Lancaster, e Steve Sparks, dell'Università di Cambridge, hanno applicato dei semplici modelli di fluidodinamica per giungere a una migliore valutazione quantitativa dei meccanismi con cui avvengono le eruzioni esplosive.

In termini generali, una eruzione è una dissipazione nell'atmosfera di energia termica contenuta nel magma. L'esplosività di una eruzione dipende dall'efficienza di trasformazione di questa energia in lavoro meccanico. Infatti una eruzione esplosiva lancia i prodotti a grande altezza ed essi si distribuiscono su una grande superficie areale. L'acqua, o un gas come l'anidride carbonica, è il tramite attraverso cui l'energia termica viene trasformata in energia meccanica. Quindi un vulcano non è altro che una macchina termica più o meno efficiente.

La quantità di acqua disciolta nel magma dipende dalla composizione chimica di quest'ultimo e dalla temperatura e pressione alle quali esso si trova. A mano a mano che un fluido magmatico risale verso la superficie diminuisce la quantità di acqua che esso può contenere in soluzione a causa della diminuizione della pressione confinante. Esiste un valore critico, chiamato «pressione di saturazione», al di sotto del quale un decremento di pressione provoca la vescicolazione, cioè l'«essoluzione» di acqua con la formazione di bolle di gas all'interno del magma. Al di sotto della pressione di saturazione, una quantità sempre maggiore di acqua essolve e contemporaneamente le bolle di gas, essendo sottoposte a una pressione confinante inferiore, si espandono e si muovono nella colonna magmatica verso l'alto. In questo modo il volume del magma aumenta e così la velocità di risalita, con una accelerazione che dipende dalla velocità con cui le bolle si espandono. L'espansione delle bolle può avvenire con modalità differenti, i cui estremi sono definiti dalla espansione adiabatica e dalla espansione isotermica. Nell'espansione adiabatica non vi è trasferimento di calore al gas che si espande, cosicché nel corso dell'espansione la sua temperatura diminuisce. Nella espansione isotermica, invece, la temperatura del gas è costante a causa di un continuo trasferimento di energia al gas. Si pensa che la vescicolazione avvenga con modalità che sono più prossime a quelle dell'estremo isotermico, perché il gas si trova a costante contatto con il magma che agisce come una sorgente di calore molto grande.

Le bolle, che si formano e si espandono, risalgono nel magma perché la loro densità è più bassa di quella del liquido. Espandendosi, le bolle vicine si uniscono formandone di nuove più grandi. Alla sommità del condotto si forma una specie di schiuma composta da magma e da bolle, nella quale il rapporto in volume fra gas e magma varia rapidamente a favore del volume occupato dal gas. L'espansione delle bolle cessa quando esse sono prossime una all'altra e la viscosità del magma impedisce ogni ulteriore aggregazione. Ciò accade quando il rapporto in volume fra gas e magma è di quattro a uno o di cinque a uno. Cessata l'espansione, la pressione del gas all'interno delle bolle aumenta rispetto alla pressione confinante. Se viene raggiunto il valore critico corrispondente alla forza

Quando avviene un'eruzione esplosiva, la risalita del magma è caratterizzata da tre stadi, illustrati a destra, i quali corrispondono a diversi gradi di trasformazione in vapore («essoluzione») dell'acqua disciolta nel magma. In profondità, il magma contenente acqua disciolta si comporta come un fluido incompressibile. Al di sopra del livello cosiddetto di essoluzione, l'acqua forma delle bolle di gas e il magma assume le proprietà di un fluido compressibile. Al di sopra del livello di frammentazione si ha una miscela costituita da gas in espansione e particelle di magma frammentato. A sinistra è indicata qualitativamente la pressione che si ha nella colonna magmatica a seconda della profondità.

La colonna che si forma nel corso di una eruzione esplosiva può avere l'andamento illustrato. Il getto alla base della colonna è costituito da una miscela turbolenta di gas e di particelle di magma alla quale si mescola l'aria atmosferica. Se la miscela risultante è più leggera dell'atmosfera circostante, si forma una colonna che sale lasciando cadere i proietti più pesanti. Quando la densità della colonna diventa uguale a quella atmosferica, la nube si espande lateralmente e viene dispersa per effetto dei venti, trascinando e facendo ricadere le particelle più pesanti.

di coesione del magma, le bolle che si trovano alla superficie cominciano a esplodere frammentandosi in pezzi minuti che rappresentano sia le pareti delle bolle («ceneri»), sia pezzi interi di magma con bolle («pomici»). Al di sopra del livello al quale avviene la frammentazione si ha una miscela di gas e di particelle fini di magma, che vengono accelerate attraverso il condotto. La profondità alla quale avviene la frammentazione è dell'ordine di qualche centinaio di metri (*si veda l'illustrazione a pagina 73*).

Se le particelle di magma che costituiscono la miscela sono di dimensioni abbastanza fini vi sarà un efficiente scambio termico tra esse e il gas che si espande isotermicamente nel condotto. La velocità che può raggiungere questa miscela è dell'ordine di varie centinaia di metri al secondo ed è funzione di differenti parametri, i più importanti dei quali sono l'alimentazione iniziale di magma, le dimensioni e la forma del condotto, e le proprietà fisiche della miscela.

In questi casi si forma un vero e proprio getto turbolento di gas e particelle che è accelerato finché la sua pressione è maggiore di quella atmosferica. La turbolenza del getto, fuori del cratere, provoca un mescolamento con l'aria, che subisce così un immediato riscaldamento. La conseguente diminuzione della densità della nuova miscela al di sotto di quella atmosferica determina l'ascensione convettiva di una nube densa di cenere. Durante alcune eruzioni esplosive si è osservato lo sviluppo di forti venti alla base del vulcano in direzione della bocca eruttiva. Questi venti, provocati dal risucchio del getto turbolento, sono a volte talmente forti che trascinano di nuovo verso il vulcano lapilli e bombe che ricadono dalla colonna eruttiva soprastante.

Il continuo mescolamento con l'aria determina una diminuzione di temperatura e infine, a una certa quota, la densità della nube e quella dell'atmosfera ritornano uguali. In questa zona si ha la massima espansione laterale della nube che viene trasportata lontano dal vulcano dai venti stratosferici.

In occasione dell'eruzione del St. Helens la quota raggiunta dalla nube è stata superiore a 27 chilometri e le particelle più fini di cenere vulcanica sono entrate in circolo nella stratosfera rimanendo sospese attorno alla Terra per vari mesi.

Il quadro che abbiamo illustrato descrive il modo in cui può avvenire una eruzione magmatica esplosiva di tipo «standard», cioè una eruzione subpliniana o pliniana. Per questo tipo di eruzione il flusso di magma che fuoriesce dal cratere è dell'ordine di varie centinaia fino a migliaia di metri cubi al secondo, l'altezza raggiunta dalla colonna eruttiva è superiore a 10 chilometri e la durata dell'eruzione è dell'ordine di ore. D'altra parte, però, le condizioni che caratterizzano una eruzione pliniana «standard» non sono facili a realizzarsi e possono intervenire fattori che modificano il modo in cui essa avviene. In particolare si possono avere condizioni che provocano il collasso della colonna eruttiva e la formazione di flussi piroclastici.

Un modello della generazione dei flussi piroclastici per collasso della colonna eruttiva è stato proposto da Sparks e collaboratori. Come abbiamo sopra accennato, affinché si realizzi una colonna convettiva deve esserci un efficiente mescolamento fra la miscela eruttiva e l'aria atmosferica. Il collasso avviene quando la densità della miscela è superiore a quella atmosferica. Questa condizione si realizza quando diminuisce la quantità di sostanze volatili nella miscela, quando diminuisce la velocità di fuoriuscita dal cratere, o quando aumenta il raggio del condotto. Una conseguenza del fenomeno è che la miscela collassante sopra il getto impedisce l'ulteriore mescolamento di quest'ultimo con l'aria circostante, cosicché, una volta iniziato, il processo tende a enfatizzarsi.

Il collasso della colonna eruttiva può però avvenire anche con un altro meccanismo proposto da Steve Malone del Geophysics Program dell'Università di Seattle e da me.

La frammentazione, che ha luogo quando la pressione all'interno delle bolle supera la forza di coesione del magma, è una vera e propria transizione di fase attraverso la quale un fluido con bolle si trasforma in una miscela di gas e particelle. Una volta iniziata, la frammentazione interessa livelli progressivamente più bassi della colonna magmatica. Allo stesso tempo, la decompressione del magma sottostante favorisce ulteriormente la nucleazione e la crescita di bolle. Se la velocità alla quale migra la superficie di frammentazione è uguale alla velocità di risalita del magma vescicolato, allora la superficie di frammentazione rimane sempre allo stesso livello e si ha una colonna sostenuta (purché le altre condizioni lo permettano). Se invece la velocità di propagazione della superficie di frammentazione è maggiore della velocità di alimentazione, allora la superficie migra verso il basso fin dove sono possibili le condizioni per la frammentazione. Poiché però le bolle richiedono un certo tempo per la loro crescita, a un certo punto la frammentazione si arresta istantaneamente con il conseguente collasso della colonna sovrastante. L'eruzione riprende con l'instaurarsi nuovamente delle condizioni iniziali che avevano permesso la frammentazione. In tal modo nel corso di una stessa eruzione possono aver luogo differenti impulsi eruttivi ciascuno dei quali è caratterizzato da una colonna sostenuta e, successivamente, da collasso e generazione di flussi piroclastici.

Tuttavia lo studio dei depositi associati con i flussi piroclastici ha fatto sospettare l'esistenza di altri meccanismi eruttivi. Infatti i depositi dei prodotti di flusso piroclastico hanno un aspetto caotico e sono generalmente sovrapposti a un deposito di materiale stratificato, il quale è formato dall'accumulo delle pomici e delle ceneri che cadono dalla colonna eruttiva prima del suo collasso. Alcuni depositi di flussi piroclastici, con strutture laminari a forma di onde, non hanno alla base alcun deposito di caduta. Dobbiamo quindi immaginare un meccanismo di generazione e di deposizione dei flussi piroclastici senza la presenza di una colonna eruttiva testimoniata dal deposito basale.

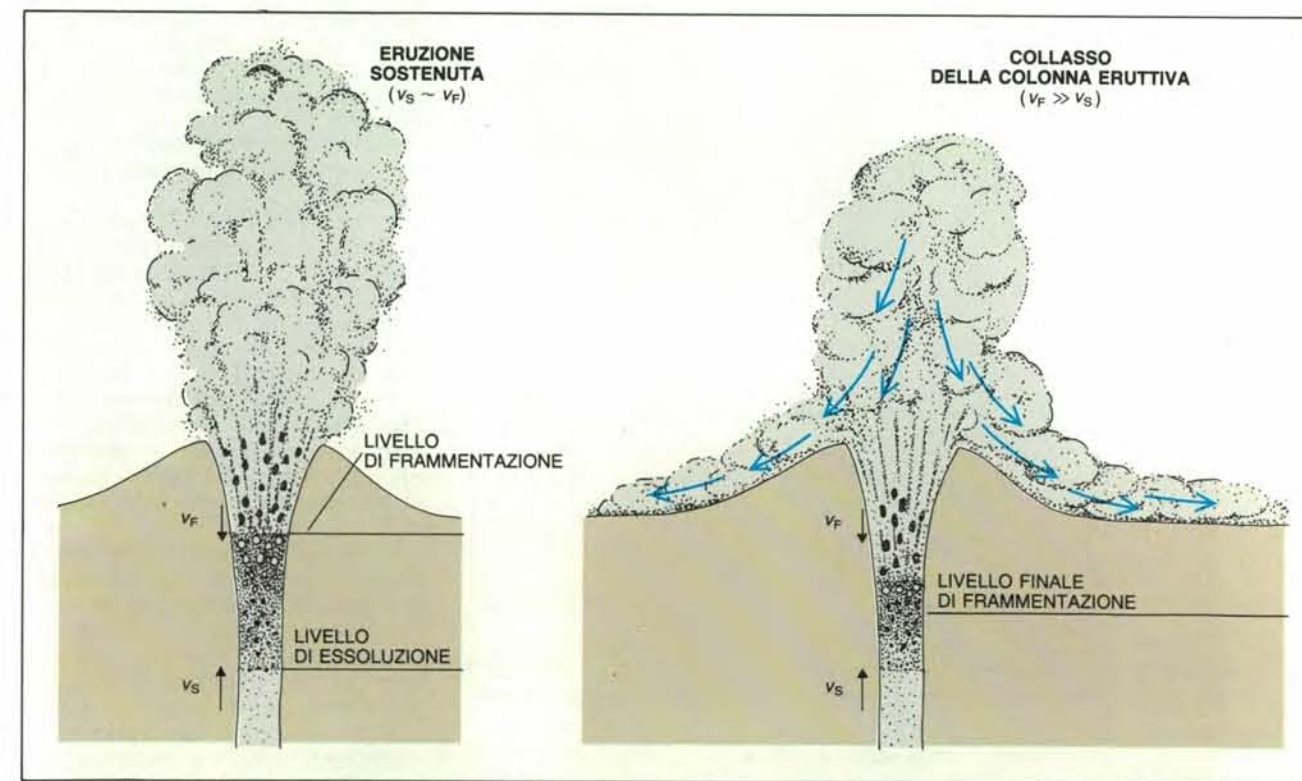
La chiave per spiegare questo mistero è legata alla pressione alla quale la miscela eruttiva fuoriesce dal cratere. Il magma risale perché esiste una differenza di pressione fra la sua zona di residenza (camera magmatica) e la superficie. Nel caso della formazione di una colonna sostenuta, la componente del moto della miscela è prevalentemente verticale perché essa giunge in superficie a una pressione pressoché simile a quella atmosferica, e tutta l'accelerazione avviene nel condotto. Se invece la miscela giunge in superficie a una pressione maggiore di quella atmosferica, allora una componente radiale dovuta alla decompressione nell'atmosfera si somma alla componente del moto verticale dovuta all'accelerazione nel condotto. Queste condizioni producono alla base della colonna eruttiva una miscela pressurizzata divergente che, una volta esaurita la fase di decompressione, scivola lungo i fianchi del vulcano per effetto della gravità.

Questo fenomeno, definito *base surge*, è stato osservato per la prima volta nelle esplosioni nucleari. In un primo tempo si è pensato che fosse connesso a una interazione fra acqua di origine esterna e magma perché ciò avrebbe giustificato le alte sovrapressioni riscontrate in superficie. Successivamente, però, si è calcolato che queste sovrapressioni possono essere generate anche in vari casi di eruzioni prevalentemente magmatiche. Si può trovare tutta una gradazione di fenomenologie che vanno dai flussi piroclastici propriamente detti fino ai base surge pressurizzati con depositi associati aventi caratteristiche abbastanza differenti.

Vale la pena di ricordare come le sovrapressioni alla bocca del vulcano possano anche essere causate dalla presenza di una massa di lava (duomo) che ostruisce il cratere. Se la sovrapressione è tale da vincere la resistenza del «tappo» solido, allora un settore di duomo può cedere facendo diminuire in maniera brusca la pressione alla quale si trova il magma sottostante.

Le eruzioni freato-magmatiche

L'interazione fra magma e acqua di origine esterna (quale per esempio quella contenuta nelle falde) perturba la normale dinamica eruttiva in quanto fornisce una grande quantità di fluido che trasformato in vapore si espande compiendo lavoro meccanico. I fattori che controllano l'evoluzione dell'interazione sono il rapporto di massa fra il magma interagente e l'acqua, la pressione alla quale avviene l'interazione, la permeabilità delle rocce, il flusso di magma, la portata della falda.

Se il magma giunge in contatto con l'acqua, una parte del suo calore viene speso per innalzare la temperatura dell'acqua e per provocarne la vaporizzazione. Questa determina un rapido aumento di volume con conseguente aumento di pressione all'interfaccia fra magma e acqua. A seconda della diffusività del mezzo, la differenza di pressione fra la zona in prossimità del magma e le zone confinanti determina un flusso di vapore e di acqua verso la zone a pres-

Differenti tipologie eruttive sono controllate dalla portata di magma durante l'eruzione. Nel caso stazionario (*a sinistra*), ossia quando la velocità di risalita (v_S) è circa uguale alla velocità di frammentazione (v_F), una colonna sostenuta si può mantenere per ore se l'alimentazione del magma è uguale alla quantità di magma effettivamente emessa. Se invece la quantità di magma emessa è maggiore (istantaneamente) di quella che giunge alla superficie di alimentazione, allora si può generare il collasso della colonna con la formazione di flussi piroclastici (*a destra*).

sione inferiore. Se il vapore ha la possibilità di diffondersi verso le rocce incassanti ne può provocare in qualche caso la frattura. Alcune volte però, in presenza di un magma vescicolato, il vapore ha la possibilità di diffondersi in esso ed essere trascinato via rapidamente; di conseguenza l'eccesso di pressione non si propaga verso le rocce.

Mike Sheridan e Ken Wohletz della Arizona State University hanno fatto interagire quantità variabili di acqua con un materiale ad alta temperatura, che simulava un magma. Essi hanno osservato che, a seconda del rapporto fra acqua e magma, l'efficienza dell'eruzione simulata varia in maniera regolare (*si veda l'illustrazione in questa pagina*). Per bassi rapporti fra acqua e magma l'attività può essere assimilata a quella stromboliana, con esplosioni discrete e lanci di materiale con traiettorie balistiche. Aumentando il rapporto aumenta anche l'efficienza della reazione, che raggiunge rapidamente il massimo quando il rapporto in peso fra acqua e magma è compreso fra 0,1 e 1. Aumentando ulteriormente il rapporto, l'efficienza diminuisce in maniera regolare, fino a che l'ulteriore aggiunta di acqua ha l'effetto di raffreddare completamente il magma.

In condizioni dinamiche, la variabile da tenere in considerazione non è tanto il rapporto fra la massa di acqua e la massa di magma quanto piuttosto i flussi di magma e di acqua che giungono in contatto.

L'effetto delle differenti condizioni è tuttora poco chiaro. Tuttavia possiamo generalmente dire che, quando la diffusione dell'impulso di pressione è verso le rocce incassanti, si può avere una esplosione causata essenzialmente dalla vaporizzazione di acqua esterna con fratturazione delle rocce incassanti e senza apprezzabile contenuto di magma; le eruzioni di questo tipo prendono il nome di «esplosioni freatiche». Quando invece si realizzano le condizioni opposte, allora si avrà un'interazione fra magma e acqua nel condotto con frammentazione accentuata del magma ed eventuale incremento delle caratteristiche esplosive; questo tipo di attività prende il nome di «attività idromagmatica».

L'eruzione del vulcano St. Helens

Dopo questo quadro generale delle fenomenologie associate alle eruzioni esplosive, siamo in grado di comprendere meglio quanto è avvenuto al Mount St. Helens nel 1980. (Per una descrizione particolareggiata dell'eruzione di questo vulcano si veda l'articolo *Le eruzioni del Mount St. Helens* di Robert Decker e Barbara Decker in «Le Scienze» n. 153, maggio 1981.) In questa sede vengono riassunti i fatti salienti in maniera da facilitare la loro interpretazione.

Alle 15 e 37 del 20 marzo 1980 i sismografi della rete sismica dell'Università di Seattle, nello Stato di Washington negli Stati Uniti nordoccidentali, registrano un terremoto di magnitudo 4,1. La stazione che per prima registra il terremoto è quella situata sul vulcano St. Helens. Il terremoto è seguito da una serie di altre scosse più piccole che avvengono con una sequenza diversa rispetto ad altre avvenute nello Stato di Washington.

Quando i sismologi dell'Università di Seattle si rendono conto che il sismografo situato sul Mount St. Helens è il primo a registrare il terremoto di magnitudo 4,1, nelle loro menti suona un campanello d'allarme. Infatti, i primi dati raccolti indicano che il terremoto è avvenuto a profondità molto superficiale e che quindi è presumibilmente di natura vulcanica. Il fatto è immediatamente notificato all'ufficio regionale del Geological Survey a Menlo Park, in California. Sul Mount St. Helens è operante un'unica stazione sismica che trasmette il segnale via radio a Seattle e, di conseguenza, la localizzazione dei terremoti è abbastanza imprecisa.

Il giorno successivo, 21 marzo, quattro nuove stazioni sismiche vengono installate in prossimità del vulcano. Nel frattempo l'attività sismica continua ad aumentare accompagnata da nuovi terremoti di magnitudo 4.

Il 23 marzo i sismologi di Seattle decidono che la sismicità può essere un segno premonitore di attività vulcanica. Tuttavia i voli di ricognizione sulla sommità del vulcano (alto circa 3000 metri) non mostrano ancora alcuna attività visibile, a parte qualche valanga causata dall'attività sismica.

Solamente il 25 marzo è possibile fotografare la formazione di una spaccatura nel manto nevoso che ricopre la cima del vulcano. Sempre in quel giorno viene interdetta all'accesso di qualsiasi visitatore la zona compresa entro un raggio di circa cinque chilometri dalla sommità del vulcano.

Il 27 marzo alle 12 e 30 si osserva la prima emissione di cenere e vapore. In quello stesso giorno viene diramata l'allerta ufficiale di rischio possibile da eruzioni vulcaniche. Da quel momento in poi numerose altre esplosioni sono registrate e studiate. L'analisi delle ceneri emesse in queste esplosioni dimostra che si tratta di esplosioni freatiche senza presenza di magma juvenile. Intanto il cratere alla sommità del vulcano continua ad allargarsi per effetto delle continue esplosioni.

L'1 aprile si nota l'allargamento del cratere e il suo abbassamento nella parte settentrionale, accompagnati da un rigonfiamento di tutto il corrispondente versante del vulcano. Alle 19 e 25 dello stesso giorno i sismografi registrano la prima comparsa di tremore armonico (cioè un segnale sismico di periodo costante causato dal movimento e dalla frammentazione del magma e che si osserva solamente nei vulcani attivi). Il numero giornaliero dei terremoti comincia a diminuire, anche se aumenta il numero di quelli di magnitudo più elevata. Nel mese di aprile l'attività continua senza grandi variazioni, ma dal confronto fra le fotografie aeree fatte nel luglio del 1979 e quelle del 12 aprile 1980 è evidente un notevole rigonfiamento del fianco settentrionale, che in alcuni punti supera gli 80 metri.

Il diverso grado di interazione fra magma e acqua può determinare differenti tipologie eruttive. Mike Sheridan e Ken Wolhetz dell'Arizona State University hanno dimostrato sperimentalmente che il massimo di efficienza meccanica nella trasformazione dell'energia termica si ha quando il rapporto in peso fra magma e acqua è compreso tra 0,1 e 1. Nell'illustrazione è mostrata la variazione della quantità di energia meccanica in funzione del rapporto fra acqua e magma.

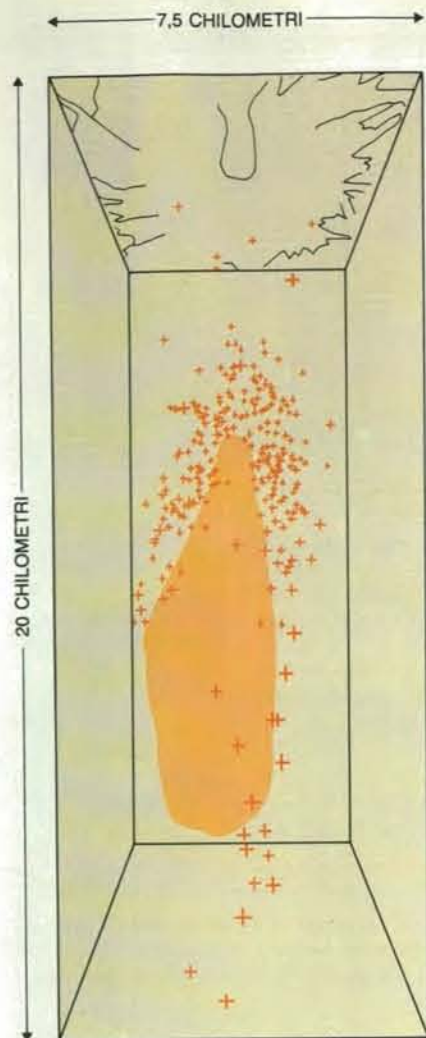
L'inizio dell'eruzione del Mount St. Helens è stato segnato da un terremoto di magnitudo 5. Nel sismogramma, di Steve Malone del Geophysics Program dell'Università di Seattle, l'inizio del terremoto è individuabile nel punto dove vi è la scritta 15 32. Più in basso si nota un periodo di relativa quiescenza sismica seguito (zona annerita) da una fase di tremore armonico che corrisponde alla fase pliniana dell'eruzione.

Il 23 aprile i geofisici del Geological Survey pongono dei riflettori sul fianco del vulcano per eseguire misurazioni distanziometriche con geodimetri a laser. In precedenza, per misurare la variazione di inclinazione del suolo, erano stati installati dei tiltmetri (livelle di grande precisione per il rilevamento di movimenti del suolo), che avevano confermato il progressivo rigonfiamento. Alla fine di aprile si ha una diminuzione del numero giornaliero di terremoti, anche se l'energia liberata rimane pressoché costante a causa dell'elevata percentuale di terremoti di magnitudo superiore a 4. Tuttavia il fianco settentrionale della montagna continua a espandersi, anche se la velocità di deformazione diminuisce e così anche il numero delle esplosioni freatiche. Nonostante questi segni, apparentemente rassicuranti, la zona «rossa» di divieto di accesso al vulcano è estesa entro un raggio compreso tra cinque e 13 chilometri. Questo provvedimento è adottato principalmente per il pericolo di frane indotte dalla instabilità dei pendii.

Alla fine di aprile il sistema di sorveglianza attorno al vulcano è costituito da 15 stazioni sismiche, cinque tiltmetri, 14 riflettori per misurazioni distanziometriche, sei stazioni per la misurazione dell'accelerazione di gravità che viene registrata due volte al giorno, tre magnetometri a lettura continua. Il Geological Survey coordina tutte queste ricerche con un ufficio a Vancouver a circa 50 chilometri dal vulcano. All'inizio del mese di maggio l'attività del vulcano continua in maniera pressoché costante. I residenti nell'area e coloro che vi lavorano premono per una riapertura della zona «rossa». I vulcanologi sanno che il vulcano può avere un'attività esplosiva con flussi piroclastici e *lahar* (colate di fango), tuttavia cercano il modo di effettuare valutazioni realistiche di rischio che garantiscano protezione alla vita umana, senza tuttavia provocare eccessivi danni economici nell'immediato futuro. Essi si sentono abbastanza sicuri di poter dare un'allerta in caso di un evento vulcanico consistente. Si basano sulla convinzione che una grossa eruzione o una frana saranno precedute da un aumento della sismicità e delle deformazioni; tuttavia non escludono la possibilità che qualcosa di anomalo possa impedire un allarme tempestivo.

Il 18 maggio il vulcanologo Dave Johnston si trova sul fianco settentrionale del vulcano per eseguire misurazioni di deformazione e analisi dei gas. Quella mattina circa un centinaio di persone si trova nell'area dichiarata inaccessibile; hanno eluso i blocchi stradali operati dalla polizia per penetrare all'interno della zona «rossa» e prendere fotografie del vulcano.

Alle 8 e 32, senza alcun altro avvertimento, un terremoto di magnitudo 5,2 scuote il vulcano. Contemporaneamente due geologi che stanno sorvolando il cratere osservano il distacco e la frana dell'intero fianco settentrionale; quando ancora la frana è in corso, essi vedono lo sviluppo di nubi di vapore ai lati della frana. Una nube di cenere si espande velocemente dal fianco settentrionale e si propaga a una velocità di circa 100 chilometri all'ora verso lo Spirit Lake. Il centro di Vancouver capta l'ultimo messaggio di Dave Johnston: «Vancouver, Vancouver... eccola...».

Tutto il fianco settentrionale del vulcano è rimosso da una esplosione latera-

Dopo l'eruzione del 18 maggio 1980, sono avvenuti sotto il Mount St. Helens vari terremoti dovuti alla rapida evacuazione di magma dalla camera magmatica. La distribuzione spaziale di alcuni di questi terremoti (*crocette*) è mostrata con la prospettiva che si avrebbe a 10 chilometri di profondità. Si può notare il volume vuoto all'interno di una nuvola di terremoti (*zona in colore*). Questo volume di dimensioni comprese tra 10 e 20 chilometri cubi è la zona dove si trova la camera magmatica.

le che devasta circa 550 chilometri quadrati di territorio. La colonna eruttiva raggiunge quote che superano i 27 chilometri. L'eruzione continua ininterrottamente per circa 10 ore con un meccanismo di tipo pliniano con l'emissione di 0,25 chilometri cubi di nuovo magma e la distruzione di 2,5 chilometri cubi di montagna. Per vari giorni, dopo l'eruzione, continuano ad avvenire terremoti sotto il vulcano.

Oramai l'equilibrio litostatico che si era formato in secoli di precedente attività è completamente sconvolto dalla distruzione del vulcano. Per questo motivo la pausa dell'attività del Mount St. Helens non è destinata a durare. Nella notte fra il 24 e il 25 maggio avviene una nuova impressionante eruzione esplosiva, preceduta da un aumento dell'attività sismica e del tremore armonico. Anche questo evento provoca un'alta colonna eruttiva e l'emissione di ceneri e pomici che dura quasi ininterrottamente per sette ore. Il magma eruttato durante questa fase è leggermente più differenziato di quello del 18 maggio: può trattarsi di magma più freddo e più vicino ai bordi della camera magmatica che ora si è avvicinato al condotto e trova una facile via di uscita verso la superficie. Anche questa eruzione è seguita da uno sciame di terremoti profondi.

Il 12 giugno, alle 19 e 05, preceduta da un periodo di ore in cui si è avuto un aumento dell'ampiezza del tremore armonico, avviene una nuova eruzione esplosiva, che dura all'incirca 10 ore ed emette un volume di materiale simile a quello del 25 maggio. Il cattivo tempo impedisce la visione del cratere, ma le osservazioni radar mostrano che dopo l'eruzione, all'interno del cratere, si è formato un duomo di lava molto viscosa. Questo duomo verrà distrutto da una nuova eruzione esplosiva che avviene il 22 luglio. Ma qualche cosa di diverso sta avvenendo nel vulcano. Dal 22 luglio le eruzioni non durano più ininterrottamente per ore, ma avvengono a brevi impulsi che durano qualche minuto e a intervalli di alcune ore. Questo è un segno inequivocabile che l'alimentazione sta diminuendo e che la velocità di frammentazione è molto più alta della velocità di alimentazione così da causare interruzioni nel corso dell'eruzione. Altre due eruzioni esplosive avvengono il 7 agosto e il 16 ottobre. Dopo l'ultima, si forma stabilmente, all'interno del cratere, un duomo di lava che viene alimentato episodicamente nel corso degli anni successivi e che attualmente si trova a una quota di circa 300 metri dalla base del cratere.

L'interpretazione della sequenza eruttiva

La distribuzione spaziale dei terremoti avvenuti sotto il Mount St. Helens, dopo le maggiori eruzioni esplosive, è peculiare. Infatti gli ipocentri si dispongono a una profondità compresa fra i sette e i 24 chilometri e circondano un volume nel quale non avvengono terremoti (*si veda l'illustrazione in questa pagina*). Si pensa che questo volume sia occupato dalla camera magmatica del vulcano, cioè dalla zona dove il magma, proveniente dal mantello, staziona e si raffredda prima di essere eruttato.

L'analisi chimica dei prodotti eruttati ha messo in evidenza che la camera magmatica era costituita, nella parte superiore, da un involucro di magma dacitico relativamente più freddo, molto viscoso e ricco di cristalli. Immediatamente sotto di esso si trovava un magma, meno ricco in cristalli ma con una percentuale di acqua in soluzione molto elevata (dell'ordine del 4-5 per cento) e, a maggiore profondità ancora, si aveva un magma progressivamente meno ricco in acqua, leggermente più povero in silice, ma con un contenuto più abbondante di cristalli.

Sulla base di questi dati si può dare una interpretazione alla successione delle varie fasi eruttive.

Il 20 marzo del 1980 un accumulo di magma relativamente poco denso, perché ricco in acqua, situato in corrispondenza del tetto della camera magmatica riesce a forzare la resistenza delle rocce sovrastanti e a spingere verso l'alto l'involucro superiore della camera magmatica relativamente più freddo e ricco in cristalli. Poiché il condotto superiore del vulcano è ostruito da un duomo di lava formatosi in precedenti eruzioni, questa massa di magma si intrude lateralmente in una zona di debolezza strutturale dell'edificio vulcanico generando terremoti. D'altra parte questa zona è anche ricca di acqua che percola dai ghiacciai e la cui improvvisa vaporizzazione causa un aumento di pressione all'interno di questo strato. Il magma che viene immediatamente a contatto con l'acqua della falda è parzialmente solido e non permette una diffusione al suo interno del vapore. Con il susseguirsi dei terremoti superficiali, questo vapore trova finalmente sfogo attraverso fratture che si propagano fino alla superficie e ha inizio così l'attività freatica. Queste esplosioni tuttavia liberano solo una parte dell'eccesso di pressione, perché nuova acqua viene fornita e trasformata cosicché l'intero fianco che poggia sull'intrusione si gonfia come un pallone. Il vulcano si trova a questo punto molto prossimo a un limite critico di instabilità tale che una piccola rottura può determinare l'esplosione dell'intero sistema. Non sappiamo che cosa possa aver causato la rottura, e neanche se il terremoto di magnitudo 5 del 18 maggio sia stato una causa o una conseguenza della frana; comunque il magma, intruso a bassa profondità e parzialmente raffreddato a causa dell'acqua di falda, dopo il franamento del fianco settentrionale si trova improvvisamente decompresso di circa 300 atmosfere, corrispondenti alla pressione esercitata dal peso della montagna. Ciò provoca la frammentazione istantanea del magma e l'esplosione fortemente direzionale.

Il processo di frammentazione comincia istantaneamente, e dura circa 12 minuti. Il riscaldamento improvviso della atmosfera, dei ghiacciai coinvolti dalla frana e delle acque del lago adiacente forma una colonna di vapore che si alza in pochi minuti a quote stratosferiche e oscura completamente il vulcano. L'esplosione abbatte, a mano a mano che si propaga, un'intera foresta di alberi secolari che, cadendo, si dispongono in direzione del flusso. Sulla base di questa direzione, Sue Kieffer dello US Geological Survey ha ricostruito la propagazione dell'onda d'urto e i parametri fisici iniziali che hanno generato l'esplosione.

Allo stesso tempo anche il magma che si trova ancora nella camera magmatica, alleggerito del carico soprastante, comincia a risalire nel condotto iniziando a vescicolare e a frammentarsi. La registrazione del tremore sismico, causato dalla frammentazione, indica che la fase propriamente pliniana dell'eruzione inizia circa alle 12 e prosegue fino a circa le 18 e 30. A questo punto una parte di magma leggermente più viscoso e forse una frana nel condotto provocano una stasi di attività. La camera magmatica, svuotata di 0,25 chilometri cubi di magma, non supporta più la pressione delle rocce confinanti e implode leggermente verso l'interno, causando i terremoti profondi. Essi avvengono a profondità maggiori di quelle a cui sono avvenuti i terremoti che hanno preceduto l'eruzione. È questa una delle prime volte in cui si osserva una migrazione dei terremoti verso il basso. Una sequenza di terremoti simile a questa verrà osservata nel 1982, durante l'eruzione del vulcano El Chichón, in Messico.

Il magma leggermente più differenziato, che aveva provocato la stasi dell'eruzione, viene emesso il 25 maggio durante una nuova fase pliniana che dura alcune ore e produce nuovi flussi piroclastici che vengono canalizzati nell'apertura a forma di ferro di cavallo che ha il nuovo cratere. Le successive eruzioni esplosive che avvengono in giugno, luglio, agosto e ottobre emettono magmi con contenuto in acqua progressivamente più basso. L'aumento progressivo di viscosità, dovuto al minore contenuto di acqua, fa diminuire la velocità di risalita del magma nel condotto e modifica il carattere delle eruzioni che da eventi a colonna sostenuta si trasformano in eventi impulsivi. L'attività di costruzione del duomo di lava, all'interno del cratere, avviene prevalentemente dal momento in cui termina l'alimentazione di magma ricco in acqua.

Al momento attuale stiamo assistendo alla ricostruzione dell'apparato vulcanico, distrutto dall'eruzione del 18 maggio, attraverso una lenta emissione di magma molto viscoso che si trova nella camera magmatica.

Il quadro offerto da questa ricostruzione si basa sui molti dati raccolti nel corso della recente attività del Mount St. Helens; si deve tuttavia sottolineare che esso è un quadro ancora parziale in quanto molti aspetti della dinamica eruttiva ci sono tuttora sconosciuti. Infatti l'inizio dell'attività eruttiva, avvenuta nel marzo del 1980, è stata correttamente prevista, ma non è stato possibile prevedere la fase parossistica del 18 maggio malgrado l'impiego di numerose strumentazioni di controllo.

Le informazioni che si raccolgono durante le eruzioni vulcaniche sono un patrimonio inestimabile per la vulcanologia e permetteranno, in futuro, una migliore comprensione dei processi fisici che avvengono all'interno del nostro inesplorato pianeta.

Scafi a propulsione umana

Nella ricerca di velocità sempre più elevate, gli scafi tradizionali, spinti da rematori, hanno dovuto cedere il passo a imbarcazioni di concezione non convenzionale, quale quella progettata e costruita dagli autori

di Alec N. Brooks, Allan V. Abbott e David Gordon Wilson

Prima d'ora gli snelli scafi da regata, o canoe da competizione, divenute familiari grazie ai Giochi olimpici, alle gare sul Tamigi e ad altre regate svoltesi in tutto il mondo erano i più veloci scafi a propulsione umana. I più veloci di questi, spinti da un equipaggio di otto vogatori, raggiungono velocità di 12 nodi su un percorso standard di 2000 metri. (Un nodo equivale a circa mezzo metro al secondo.) Oggi, gli scafi a propulsione umana che non sono vincolati dai regolamenti ufficiali di queste gare sono in grado di superare queste velocità. I progettisti di scafi non convenzionali stanno scartando i remi a favore di eliche ad alto rendimento; non solo, stanno anche facendo a meno delle carene, esplorando nuove vie per ridurre la resistenza all'avanzamento, ossia la resistenza idrodinamica, che l'acqua esercita sullo scafo in movimento.

In verità, due di noi (Brooks e Abbott) hanno sviluppato per la conquista di un record uno scafo a propulsione umana: il *Flying Fish II*, che si guida come una bicicletta. Fornito di una coppia di ali idrodinamiche, o ali sommerse, e di un'elica ad alto rendimento, questo scafo permette a un unico pilota di completare un percorso di 2000 metri in modo sensibilmente più veloce di quanto faccia un singolo rematore su uno scafo da competizione e ha raggiunto la velocità massima di 13 nodi su percorsi brevi.

A prescindere dal particolare progetto - che sia un rudimentale oggetto galleggiante che avanza grazie a spinte subacquee, una zattera di legno spinta da pali, un tronco scavato mosso da pagaie o un piccolo scafo munito di remi convenzionali - ogni scafo deve contendere con quattro forze di base: peso, spinta idrostatica, spinta propulsiva e resistenza idrodinamica (senza considerare in questa sede la resistenza aerodinamica). Peso e spinta idrostatica sono le forze più facili da comprendere. Il peso è semplicemente la forza gravitazionale che tira verso il basso lo scafo e i suoi occupanti. La spinta idrostatica è la forza che agisce verso l'alto, controbilanciando il peso. Fino a che un'imbarcazione non viene sottoposta ad alcuna accelerazione verticale, la spinta idrostatica è uguale in modulo al peso.

Per la maggior parte degli scafi, la spinta idrostatica è dovuta alla spinta di galleggiamento, ossia allo spostamento dell'acqua provocato dalla carena dello scafo. La spinta idrostatica corrisponde in modulo al peso del volume dell'acqua spostata e questo vale anche in assenza di movimento. Inoltre, molti scafi da alta velocità traggono vantaggio dal sostentamento dinamico che è prodotto dal movimento dello scafo nell'acqua. Un esempio comune di sostentamento dinamico è dato dalla planata: quando il fondo della carena devia in continuazione l'acqua verso il basso si produce un sostentamento come forza di reazione. Una imbarcazione che fruisca di un elevato sostentamento dinamico durante la planata si muove con la maggior parte dello scafo fuori dall'acqua e richiede una minore spinta di galleggiamento. Fino a poco tempo fa i progettisti di scafi a propulsione umana non erano in grado di ottenere un sostentamento dinamico dei loro scafi.

La spinta propulsiva è la forza (prodotta dall'azione dell'uomo nel caso degli scafi a propulsione umana) che muove lo scafo. La resistenza idrodinamica è la forza che per definizione agisce in direzione opposta all'avanzamento dello scafo. Se una imbarcazione, raggiunta una certa velocità, si muove di moto costante, la spinta propulsiva è uguale in modulo alla resistenza idrodinamica. In sostanza, a velocità costante la spinta idrostatica bilancia il peso e la spinta propulsiva bilancia la resistenza.

Per trasformare efficientemente una data energia umana d'ingresso in velocità è necessario innanzitutto ridurre al minimo la resistenza idrodinamica. Un sistema ovvio è quello di ridurre il peso dell'imbarcazione. Quando questa comincia a muoversi la sua fonte di spinta idrostatica (cioè il volume della carena) quasi sempre comporta una penalizzazione in resistenza idrodinamica. Riducendo al minimo il peso dell'imbarcazione, si riduce la spinta idrodinamica richiesta e quindi la resistenza idrodinamica connessa alla spinta idrostatica è inferiore. Partendo dal presupposto che i membri dell'equipaggio non siano in sovrappeso, la riduzione di peso deve riguardare innanzitutto lo scafo.

Sforzi in questo senso hanno prodotto scafi da competizione che pesano quanto una piccola frazione del peso del rematore, con un rapporto paragonabile a quello fra una moderna bicicletta da corsa e il ciclista. Nel passato gli scafi erano di solito costruiti in legno di cedro, abete e mogano ed erano resi più leggeri assottigliando lo spessore delle loro carene. (Infatti il termine «shell» - guscio - derivava proprio dal fatto che una leggera pressione del dito poteva facilmente provocare un buco nel legno della carena.) Negli anni cinquanta vennero sperimentati alcuni scafi che avevano una sottile carena di vetroresina rinforzata e dalla fine degli anni sessanta vennero introdotte imbarcazioni realizzate in materiali compositi che entrarono in aperta con-

Le imbarcazioni tradizionali da regata e quella non convenzionale progettata dagli autori, il *Flying Fish II*, sono tra i più veloci scafi a propulsione umana. L'imbarcazione a bordo della quale sta remando il quattro volte olimpionico John Van Blom è stata costruita dallo svizzero Alfred Stämpfli. *Flying Fish II*, con a bordo uno degli autori (Abbott), è spinto da un'elica mossa a pedali ed è sostenuto da due ali idrodinamiche, o ali sommerse. I galleggianti durante la planata sono fuori dall'acqua poiché sono stati ideati solo per sostenere lo scafo alle basse velocità.

correnza, negli ambienti agonistici, con gli scafi di legno, fino ad allora predominanti. Oggi uno scafo di legno è una rarità. Sofisticati materiali compositi, consistenti in una matrice di resina unita a fibre polimeriche o a grafite, hanno significativamente abbassato il peso della più leggera imbarcazione monoposto al di sotto dei 10 chilogrammi.

Dal momento che il peso di uno scafo da competizione poteva così considerarsi ridotto al minimo, l'attenzione del progettista ha dovuto rivolgersi altrove per ridurre la resistenza idrodinamica. Gli scafi hanno ciò che viene comunemente indicato come carena dislocante: tutta la loro spinta di sostentamento è prodotta dalla spinta di galleggiamento della carena. Le carene dislocanti hanno come unica proprietà quella che la loro resistenza idrodinamica tende a zero con il tendere a zero della loro velocità rispetto all'acqua. Di conseguenza a velocità molto basse gli scafi a carena dislocante incontrano una resistenza idrodinamica molto bassa e sono fra i più efficienti di tutti gli scafi. Le imbarcazioni da competizione, comunque, non sono fatte per le basse velocità.

A mano a mano che la velocità di uno scafo aumenta, cresce in modo evidente anche la sua resistenza idrodinamica a causa, almeno in parte, della formazione di onde prodotte dalla prua e dalla poppa. L'energia necessaria per produrre queste onde si manifesta come resistenza d'onda. Al crescere della velocità la resistenza d'onda cresce rapidamente, ma in modo irregolare poiché l'onda di prua può interagire con quella di poppa positivamente (quando le onde sono in fase e si rinforzano reciprocamente) o negativamente (quando le onde sono fuori fase e tendono ad annullarsi l'una con l'altra). A una velocità nota come velocità di carena, la prua è sulla cresta e la poppa è nell'avvallamento di un'unica onda; nel suo passaggio attraverso l'acqua, la carena ha letteralmente formato una collina d'acqua attraverso la quale deve essere spinta l'imbarcazione. A questo punto è necessario un grande dispendio di energia per aumentare la velocità del mezzo, ma la «macchina» uomo non può sopperire allo sforzo richiesto sicché la velocità di carena si comporta come se fosse il limite effettivo della velocità di uno scafo con carena dislocante spinto da propulsione umana.

La velocità di carena di un'imbarcazione è proporzionale alla radice quadrata della sua lunghezza alla linea di galleggiamento. Imbarcazioni a propulsione umana che abbiano lunghe carene dislocanti sono quindi meno penalizzate dalla resistenza d'onda di quanto lo siano, a parità di spinta di galleggiamento totale, imbarcazioni che hanno carene corte. D'altro canto, per un data spinta di galleggiamento le carene snelle hanno maggior superficie bagnata di quanta ne abbiano le carene corte e larghe. Quanto più grande è la superficie bagnata, tanto maggiore è la resistenza idrodinamica causata dall'attrito dovuto all'acqua che scorre sulla superficie della carena. Questo tipo di resistenza idrodinamica è conosciuto come resistenza di attrito (della superficie bagnata). Quindi, se a un'imbarcazione viene data una forma più snella, la resistenza d'onda diminuisce, ma la resistenza di attrito diventa un problema maggiore.

A una carena progettata per imbarcazioni veloci deve quindi essere data una forma tale da ridurre al minimo la somma della resistenza d'onda e della resistenza di attrito. Sono state progettate carene per compiere percorsi di sei-sette minuti alla potenza di circa mezzo cavallo vapore per rematore. (Un cavallo vapore è uguale più o meno a 750 watt.) Il rapporto ottimale lunghezza-larghezza trovato per queste snelle imbarcazioni supera il valore 30. Uno scafo con un solo rematore, per esempio, ha una lunghezza compresa fra gli otto e i nove metri, mentre la larghezza non supera i 30 centimetri. Di conseguenza, la forma di scafo ottimale presenta una distribuzione dissimmetrica della resistenza idrodinamica alle velocità da competizione: l'80 per cento della resistenza che agisce sullo scafo è dovuta alla resistenza di attrito e il 20 per cento è dovuto alla resistenza d'onda.

Dato che la resistenza di attrito rappresenta la fonte principale della resistenza idrodinamica che agisce sullo scafo alle alte velocità, una riduzione sostanziale di quest'ultima è possibile se può essere ridotta la resistenza di attrito. Questa prende origine da un sottile velo d'acqua, conosciuto come strato limite, che scorre lungo la carena dell'imbarcazione. Esistono due tipi fondamentali di strato limite: uno laminare, in cui lo scorrimento è lineare e stazionario, e uno turbolento, in cui lo scorrimento è caotico e non stazionario. Gli strati limite laminari producono una resistenza di attrito molto inferiore a quella prodotta dagli strati limite turbolenti. Lo strato limite su uno scafo è laminare a prua, ma spostandosi anche di poco verso poppa si trasforma in un flusso turbolento. La resistenza idrodinamica è sensibilmente ridotta se questa transizione è ritardata, aumentando così l'area dello scorrimento laminare sulla carena.

Uno dei sistemi per aumentare lo strato limite laminare, applicato su alcuni particolari mezzi subacquei, è l'iniezione di polimeri a catena lunga nello strato limite vicino alla prua dell'imbarcazione. Le organizzazioni che stilano i regolamenti delle gare non sono molto favorevoli ad ammettere questo sistema nelle competizioni, se non altro perché inquina l'acqua. Un procedimento maggiormente accettabile sarebbe quello di coltivare con molta cura uno strato di alghe o altri innocui microrganismi per creare un velo viscido sulla carena.

L'aspirazione dello strato limite rappresenta un'altra tecnica adottata per stabilizzare lo strato limite laminare. In

SPINTA IDROSTATICA

RESISTENZA IDRODINAMICA

PESO

SPINTA PROPULSIVA

Nella progettazione di un'imbarcazione simile a questa si devono tenere in considerazione quattro forze basilari: peso, spinta idrostatica, spinta propulsiva e resistenza idrodinamica. Il peso è la forza gravitazionale che agisce sullo scafo e sul pilota. La spinta idrostatica, detta anche comunemente spinta di Archimede, è la forza rivolta verso l'alto pari al peso dell'acqua spostata dalla carena. Un'ulteriore spinta verso l'alto, il sostentamento dinamico, può essere prodotta dal fluire dell'acqua sotto la carena. La spinta propulsiva, nel caso di natanti mossi da energia umana, è la forza prodotta dai movimenti del pilota che spinge avanti l'imbarcazione. La resistenza idrodinamica, la forza che si oppone all'avanzamento dell'imbarcazione, deriva, nella maggior parte degli scafi, dalla formazione di una scia (resistenza d'onda) e dall'attrito fra la carena e l'acqua (resistenza di attrito della superficie bagnata). Per mantenere costante la velocità di una imbarcazione, la spinta idrostatica deve bilanciare il peso e la spinta propulsiva deve bilanciare la resistenza idrodinamica. Obiettivo chiave nella progettazione di un'imbarcazione è la riduzione al minimo della resistenza idrodinamica alla normale velocità operativa. Per raggiungere velocità da competizione, si riduce la resistenza con il ricorso a scafi leggeri, lunghi e stretti.

La potenza necessaria a un'imbarcazione a propulsione umana per raggiungere una determinata velocità per un certo periodo di tempo dipende dal progetto dello scafo. Il grafico a sinistra mostra la potenza richiesta per uno scafo a remi, che si affida al dislocamento dell'acqua prodotto dalla sua carena per la maggior parte del suo sostentamento, paragonato con altri due progetti di scafi (uno progettato e costruito da due degli autori) che si affidano al sostentamento dinamico prodotto da ali idrodinamiche. Alle basse velocità gli scafi a dislocamento sono più efficienti degli scafi con ali idrodinamiche. Attualmente gli scafi con ali idrodinamiche hanno una velocità minima al di sotto della quale le ali idrodinamiche non permettono di sostenere il peso totale dello scafo e del pilota. A velocità più elevate, tuttavia, la situazione muta: le imbarcazioni ad ali idrodinamiche sono più efficienti degli scafi a dislocamento. Il grafico a destra mostra come la potenza che un atleta può fornire diminuisce con la durata dello sforzo. Per periodi limitati la potenza generata dal movimento circolare dei pedali è notevolmente più alta di quella generata remando. Un progetto ottimale con ali idrodinamiche potrebbe rendere possibile il raggiungimento di velocità superiori ai 20 nodi. Una tale impresa richiederebbe livelli di potenza che possono essere ottenuti solamente pedalando e solo per pochi secondi. Uno scafo che montasse una tale ala idrodinamica avrebbe però difficoltà a partire: la sua velocità di «decollo» dovrebbe superare gli 11 nodi.

questo procedimento il fluido nello strato limite è continuamente «succhiato» all'interno dell'imbarcazione attraverso pori o piccole fessure sulla superficie della carena. La tecnica dell'aspirazione dello strato limite potrebbe essere effettivamente impiegata sugli scafi se questi venissero costruiti con carene in materiali porosi che permettessero all'acqua di filtrare capillarmente. Potrebbe venir utilizzata una piccola pompa per eliminare, di tanto in tanto, l'acqua che viene aspirata.

La rifinitura della superficie bagnata della carena può anche influire sulla riduzione della resistenza di attrito. Studi condotti sotto l'egida della National Aeronautics and Space Administration hanno dimostrato che una superficie lisciata a cera non sempre comporta una diminuzione della resistenza di attrito. Superfici con sottilissime scanalature nella direzione del flusso, dette «*riblet*», hanno presentato il sei per cento in meno di resistenza di attrito rispetto alle superfici completamente lisce.

Queste scanalature sono state provate da un equipaggio di rematori provenienti da un gruppo del Flight Research Institute diretto da Douglas McLean della Boeing Company. Uno scafo monoposto venne ricoperto con un rivestimento sperimentale di plastica in cui erano state praticate delle scanalature. Lo spazio fra le scanalature era la tremillesima parte di un pollice (circa 80 micrometri), inferiore quindi all'intervallo fra i solchi di un disco fonografico. Le prove dimostrarono che la velocità massima dello scafo era maggiore del due per cento. Sebbene questo valore possa sembrare insignificante, su un percorso standard di 2000 metri è equivalente a un vantaggio di quattro lunghezze.

Sulla base di questi incoraggianti risultati, il rivestimento sperimentale venne applicato sulla carena dello scafo di un equipaggio olimpico americano «quattro con». (Un'imbarcazione «con» è governata da un timoniere che non rema, ma che dà la cadenza della remata all'equipaggio.) L'equipaggio raggiunse un ottimo risultato, conquistando la medaglia d'argento nei giochi estivi del 1984.

In aggiunta alla bassa resistenza idrodinamica, un altro ingrediente essenziale per portare un'imbarcazione al successo è un buon rendimento nella propulsione: quanta più energia umana possibile deve essere convertita in spinta propulsiva utile. Nel caso della remata, i due maggiori progressi nel rendimento della propulsione risalgono alla metà del XIX secolo. Uno fu lo sviluppo del moderno supporto per lo scalmo nel 1843. Il supporto è un congegno attaccato al fianco dell'imbarcazione: lo scalmo (o punto di fulcro) per il remo si trova all'estremità del supporto a treppiede. Poiché non era più necessario attaccare lo scalmo direttamente sulla cinta dello scafo o sul lato, la carena poteva essere più stretta (riducendo la resistenza idrodinamica d'onda) e i remi potevano essere più lunghi (permettendo ai rematori di dare remate più lunghe ed efficaci).

Il secondo progresso fu realizzato nel 1856: il seggiolino scorrevole. Fino a quel momento le imbarcazioni a remi erano spinte dallo sforzo dei muscoli delle braccia, delle spalle e della schiena, mentre i potenti muscoli delle gambe venivano utilizzati soltanto per tenere in allenamento il corpo o per fargli da sostegno. La remata comportava un notevole sforzo per vincere lentamente la resistenza. L'installazione del seggiolino scorrevole fece sì che l'energia dei muscoli delle gambe venisse utilizzata, poiché il sedile si muove avanti e indietro quando si flettono e si allungano le gambe durante il ciclo completo della remata. Il primo seggiolino scorrevole era piuttosto rudimentale e consisteva in un cuscino di pelle di capra che scivolava su un pannello lubrificato con del grasso. Il sedile su cuscinetti, ancora in uso oggi, fu inventato negli Stati Uniti nel 1857.

Nel 1981 su uno scafo con una variante del sedile scorrevole Peter Michael Kolbe vinse il campionato del mondo a Monaco. A differenza degli scafi convenzionali che hanno sedili scorrevoli, ma supporti e traversine, o pedane, fissi, l'imbarcazione realizzata per Kolbe era dotata di un sedile fisso e di una struttura scorrevole che sosteneva le traversine (e gli scalmi). Con questa combinazione, il movimento per la remata è uguale a quello che si farebbe su uno scafo con-

venzionale, ma dal momento che la maggior parte della massa del rematore grava sul sedile fisso (e non su quello scorrevole), le oscillazioni del centro di massa (che più o meno coincide con il rematore) sono notevolmente ridotte. Ciò, a sua volta, fa diminuire le oscillazioni di velocità a cui uno scafo è soggetto quando si muove attraverso l'acqua. (In uno scafo convenzionale queste oscillazioni si manifestano, quando si rema con una certa foga, con il caratteristico avanzamento a scatti.)

Poiché la resistenza di attrito non è una funzione lineare della velocità dell'acqua rispetto alla carena (infatti è proporzionale al quadrato della velocità), una velocità variabile produce sempre più resistenza idrodinamica di quanto avverrebbe se l'imbarcazione si muovesse in modo costante alla velocità media. La riduzione della resistenza idrodinamica delle appendici ottenuta con l'impiego dello scalmo scorrevole è scarsa, ma sufficiente a creare una significativa differenza in gara. Nella finale del campionato del mondo del 1982 cinque imbarcazioni avevano sedili fissi e supporti degli scalmi scorrevoli. Nel 1983 tutti e sei i finalisti gareggiavano su imbarcazioni con supporti scorrevoli. Dopo il 1983, tuttavia, questo tipo di imbarcazione venne escluso dalle competizioni.

Benché l'aggiunta del supporto degli scalmi e del sedile scorrevole avesse aumentato sensibilmente il rendimento propulsivo della remata, la remata stessa aveva una limitazione fondamentale. Remi e pagaie sono di per se stessi congegni di attrito: producono una spinta propulsiva slittando indietro attraverso l'acqua. Lo slittamento rappresenta una perdita di rendimento che può essere limitata aumentando la superficie della pala del remo, ma solo entro certi limiti, per motivi pratici. Per di più la resistenza aerodinamica causata dalle pale quando sono fuori dall'acqua, durante la fase di ritorno, può essere elevata, in particolar modo in presenza di vento forte.

Il rendimento di un sistema propulsivo è definito come rapporto fra potenza utile prodotta, che è il prodotto della spinta propulsiva media per la velocità, e la potenza umana applicata. Non tutti i dettagli fisici, per quanto concerne la remata, sono conosciuti, ma studi fatti da diversi ricercatori hanno collocato il rendimento propulsivo della remata fra il 65 e il 75 per cento. Quindi circa due terzi della potenza prodotta da un rematore vengono trasmessi all'imbarcazione come lavoro utile; il resto va a creare perturbazioni nell'acqua e nell'aria.

Nelle imbarcazioni a motore sono stati abbandonati molto tempo fa i sistemi di propulsione con congegni di attrito, come la ruota, a favore dell'elica. È interessante notare come, prima dello sviluppo dei piccoli fuoribordo a benzina all'inizio del secolo, i mezzi a propulsione umana dotati di elica fossero in fase di sviluppo nel settore dei mezzi di trasporto. Tali scafi spinti da eliche si sono dimostrati maggiormente veloci e meno stancanti di canoe o di imbarcazioni a remi. Nel 1890 un catamarano (un'imbarcazione con due scafi), spinto a eliche da un equipaggio di tre persone, si dimostrò più veloce del 13 per cento rispetto a uno scafo con tre rematori su un percorso di 163 chilometri lungo il Tamigi.

Le eliche presentano alcuni svantaggi: possono imbrattarsi con alghe e possono toccare il fondo nei bassi fondali; tuttavia sono particolarmente adatte alla propulsione umana. Si possono sfruttare progetti di eliche ad alto rendimento le cui pale abbiano un grado di snellezza maggiore, poiché il livello di potenza è abbastanza basso. Per di più, le velocità periferiche sono sufficientemente basse per non creare problemi di cavitazione. (La cavitazione è la formazione di bolle di vapore quando la pressione assoluta su alcune parti dell'elica che gira è ridotta al di sotto della tensione di vapore dell'acqua; il fenomeno riduce il rendimento e può causare un'eccessiva corrosione sulla superficie delle pale.) Parecchi nuovi scafi a propulsione umana sono stati dotati di eliche il cui rendimento supera il 90 per cento.

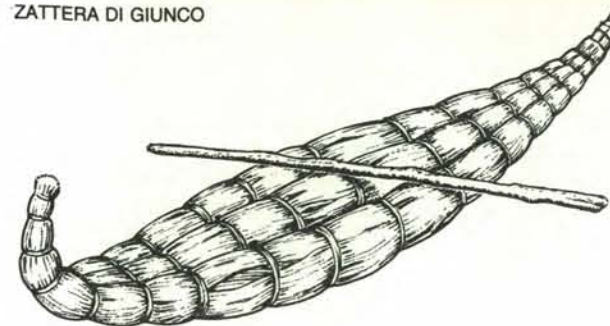
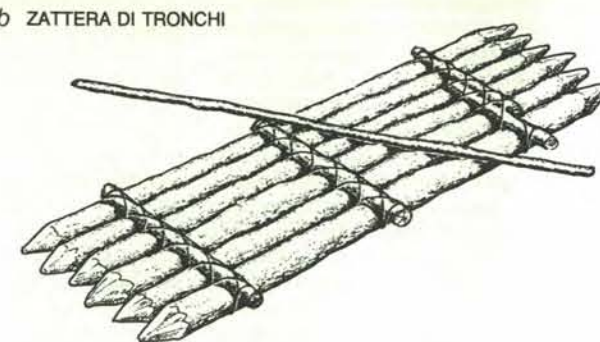
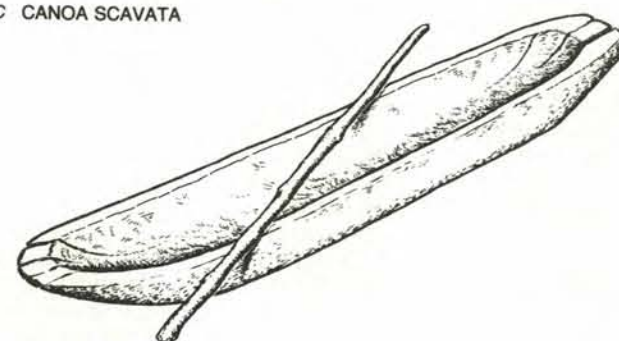
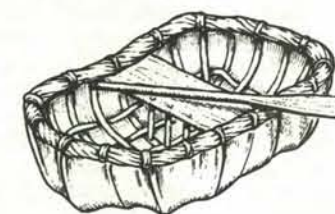
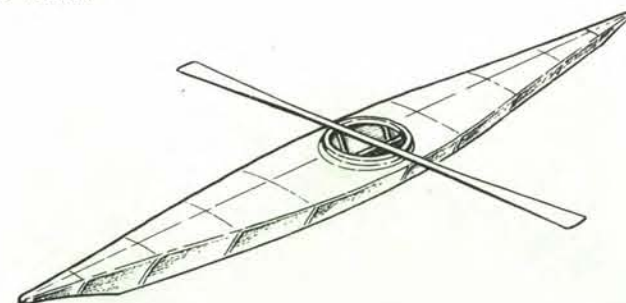
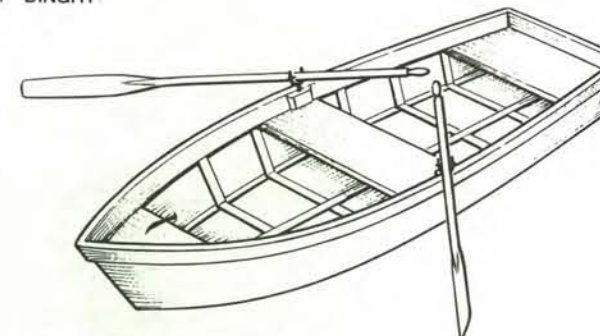
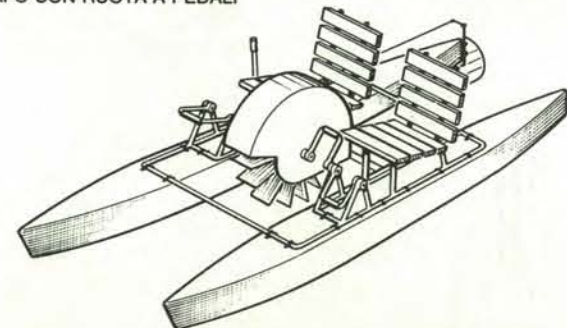
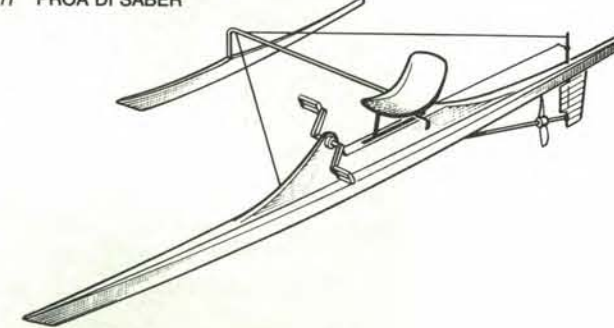
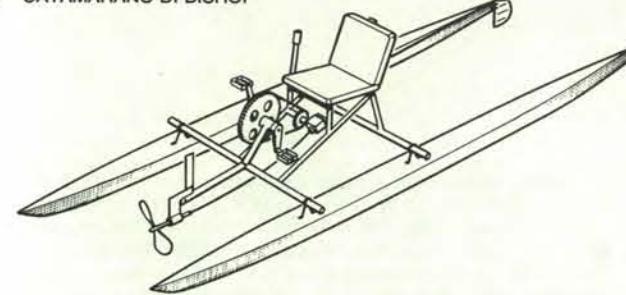
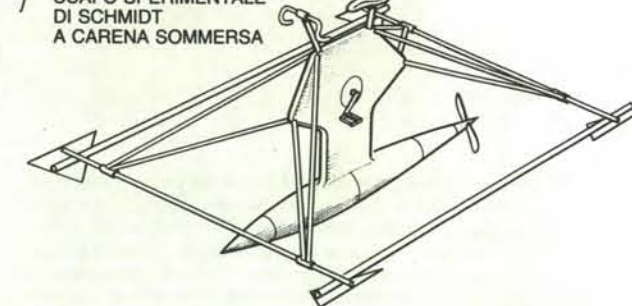
Il moto rotatorio di un'elica è relativamente facile da ottenere con un sistema di pedali, ingranaggi e catene molto simili a quelli di una bicicletta. Un sistema del genere trae vantaggio dai forti e rapidi movimenti delle gambe. Il movimento circolare del pedalare su una bicicletta rimane il più efficiente e pratico sistema per un trasferimento continuo di energia da un corpo umano a una macchina. (Non è una coincidenza che i mezzi a propulsione umana sia aerei sia su terra con cui sono stati stabiliti record si basino sul sistema di trasmissione a catena di bicicletta.)

Un campione di ciclismo può produrre quasi due cavalli di potenza per alcuni secondi di massimo sforzo. Per periodi di sforzo continuo oltre i sei minuti, invece, la potenza prodotta non supera il mezzo cavallo. Parecchi fattori influiscono sulla produzione di potenza: l'andamento della pedalata, l'altezza della sella, la lunghezza del braccio del pedale, le condizioni fisiche e la determinazione del ciclista. Il movimento tradizionale del remare, dove il rematore è seduto fermo e aziona soltanto i muscoli della schiena, delle braccia e delle spalle, produce considerevolmente meno potenza di quanto faccia il movimento del pedalare del ciclista. L'introduzione del seggiolino scorrevole ha incrementato la potenza della remata fino a rivaleggiare con quella della pedalata - almeno per periodi superiori ai pochi minuti. (Il vantaggio viene perso però dopo circa un minuto, per i limiti posti dai sistemi circolatorio e respiratorio dell'uomo.)

I vantaggi relativi a uno scafo spinto a pedali rispetto a quelli di uno scafo a remi stanno quindi in prevalenza nel fatto che i remi sono meccanismi meno efficienti per convogliare energia umana nella propulsione di uno scafo. Per di più la velocità discontinua di un'imbarcazione a remi, rispetto all'andatura senza strappi che si può raggiungere con un'elica, determina una maggiore penalizzazione nella resistenza idrodinamica.

Alcuni progettisti di scafi veloci a propulsione umana hanno anche tentato di ridurre la resistenza idrodinamica con sistemi nuovi. Un modo per eliminare potenzialmente la resistenza d'onda e nel medesimo tempo ridurre la resistenza di attrito è quello di sommergere la carena; in questo caso la persona a bordo dovrebbe trovarsi fuori dall'acqua sostenuta da stretti supporti sporgenti sopra la carena. Una carena con la minima resistenza idrodinamica ha, in questo caso, la forma di una goccia d'acqua con una lunghezza compresa fra tre e quattro volte la sua larghezza.

Per il pilota sarebbe difficile, se non impossibile, tenere in equilibrio questo tipo di struttura, paragonabile a quella di un monociclo. Theodore Schmidt ha superato parzialmente questo problema collegando quattro piccole ali idrodinamiche a una carena sommersa sperimentale da lui stesso progettata. Una geometria a triciclo con tre carene sommerse più piccole sarebbe più stabile ma non altrettanto efficiente. Dacché il rapporto tra superficie dello scafo e dislocamento diventa più piccolo a mano a mano che il dislocamento aumenta, una carena grande ha un'area superficiale minore di tre carene piccole con la stessa spinta di galleggiamento totale.

Il problema dell'equilibrio di una carena singola subacquea potrebbe essere risolto collocando il guidatore nella carena, creando così un sommergibile. Ma una carena a profilo aerodinamico, abbastanza grande per contenere una persona, dislocherebbe una quantità di ac-

Gli scafi a propulsione umana possono differire per forma, per materiali di costruzione e per sistemi di propulsione. Gli scafi relativamente primitivi erano spinti da pali (*a-c*) o da pagaie (*d,e*) e costruiti con i più svariati materiali naturali, come giunco, legno e pelli di animali. Gli scafi moderni sono costruiti in legno o in metallo e spinti da remi (*f*), un sistema che richiede l'uso delle braccia, delle spalle, della schiena, o sono spinti da ruote a pedali, mettendo in azione i forti muscoli delle gambe (*g*). Gli scafi a elica spinti da pedali (*h-j*) sono più efficienti delle barche a remi o con ruota a pedali. Progetti e materiali nuovi hanno anche ridotto la resistenza idrodinamica che questi scafi incontrano. La carena sommersa dello scafo sperimentale di Theodore Schmidt (*j*) elimina con successo, per esempio, il problema della resistenza d'onda.

qua molto maggiore e avrebbe una area superficiale superiore a quella di una carena con una spinta di galleggiamento giusto sufficiente a sostenere il peso di una persona. Sebbene un sommergibile a propulsione umana non sia ottimale come mezzo di trasporto vicino alla superficie dell'acqua, esso potrebbe costituire un notevole miglioramento rispetto a un sommozzatore con pinne. All'inizio degli anni cinquanta un sommergibile di questo tipo spinto da due persone, progettato da Calvin Gongwer e chiamato Mini-Sub, venne prodotto in un numero ridotto di esemplari dalla Aerojet-General Corporation. Spinto da una coppia di eliche controrotanti di 760 millimetri di diametro, pare che il Mini-Sub potesse raggiungere sette nodi di velocità: circa tre volte la velocità alla quale un sommozzatore può nuotare sott'acqua.

Altri progetti cercano di ridurre la seconda maggior componente della resistenza idrodinamica, la resistenza di attrito, sfruttando il sostentamento dinamico per far uscire dall'acqua parte dello scafo così da ridurre la superficie bagnata. Benché il sostentamento dinamico provochi una penalizzazione in resistenza idrodinamica, in tanti casi la riduzione della resistenza di attrito riesce più che a compensare la resistenza idrodinamica dovuta al sostentamento dinamico.

Scafi a propulsione umana che usano il sostentamento dinamico per planare sono ancora nella fantasia dei progettisti, ma è stato usato con successo un altro modo per generare il sostentamento dinamico: le ali idrodinamiche. Le ali idrodinamiche sono ali che funzionano sott'acqua producendo un sostentamento, analogamente all'azione delle ali di un aeroplano. La dimensione di un'ala idrodinamica è abbastanza modesta in paragone a quella di un'ala di aeroplano. Per esempio, alla velocità di nove nodi è necessario, per produrre abbastanza sostentamento da reggere una persona fuori dall'acqua, qualcosa di meno della decima parte di un metro quadrato di superficie alare. Un'ala idrodinamica progettata per produrre il medesimo sostentamento a una velocità doppia richiederebbe solo un quarto della superficie precedente.

Sebbene la ridotta superficie bagnata di un'ala idrodinamica provochi una minima resistenza di attrito, le ali idrodinamiche danno origine a un altro tipo di resistenza idrodinamica. L'ala idrodinamica si muove attraverso l'acqua lasciando dietro di sé una scia di vortici analogamente a quanto avviene con le ali di un aeroplano. L'energia spesa per generare la scia di vortici si manifesta come resistenza indotta. Anche lo spruzzo sollevato dai supporti verticali dell'ala idrodinamica, a mano a mano che tagliano la superficie dell'acqua, rappresenta una resistenza supplementare.

Un altro grande problema della propulsione umana con mezzi ad ali idrodinamiche è quello di raggiungere nell'acqua una velocità relativamente elevata necessaria per potere decollare e sostenersi sull'acqua. Siccome a velocità zero l'ala idrodinamica non produce sostentamento, è necessario un altro sistema di galleggiamento, come una carena dislocante, all'inizio e alla fine del «volo». Un'ala studiata per planare sull'acqua dovrebbe muoversi a 10 nodi prima di generare una spinta di sostentamento tale da reggere lo scafo e il pilota. Questa velocità potrebbe essere impossibile da raggiungere mentre lo scafo è ancora sostenuto sull'acqua dalla sua carena a dislocamento. Un'ala idrodinamica più grande ridurrebbe la velocità di decollo, ma la resistenza provocata dall'incremento di superficie non permetterebbe allo scafo di andare così veloce.

Flying Fish II è uno scafo a propulsione umana con ali idrodinamiche, progettato e costruito da due degli autori (Brooks e Abbott). Lo scafo, spinto da un congegno a pedali, con elica ad alto rendimento, decolla a sei nodi e ha una velocità massima di circa 14 nodi. Si guida proprio come una bicicletta. La prima versione non aveva galleggianti laterali e aveva quindi bisogno di una rampa di lancio a catapulta per raggiungere la velocità di decollo. Nell'attuale versione i galleggianti sostengono lo scafo in modo tale che si possa raggiungere la velocità di decollo con partenza da fermo. La profondità a cui le ali idrodinamiche «volano» è controllata automaticamente da una leva a forma di spatola che segue la superficie dell'acqua ed è collegata a un sottile flap (timone di profondità) sull'ala idrodinamica frontale. Lo scafo ha coperto un percorso di 2000 metri in un tempo di circa 10 secondi inferiore a quanto sia il record di un rematore a bordo di uno scafo da competizione.

Sovrapponendo una sull'altra le ali idrodinamiche in modo tale che le più piccole siano sotto alle più grandi, come avviene negli aliscafi a motore, si potrebbe aggirare il problema. Uno scafo con ali a scala, di larghezza decrescente, potrebbe decollare alle basse velocità sulle ali superiori più grandi. Una volta raggiunta una velocità abbastanza elevata perché il sostentamento prodotto dalle ali inferiori sia sufficiente per sostenere lo scafo, questo si alzerebbe ulteriormente togliendo dall'acqua le ali più grandi e riducendo così la resistenza. A causa delle difficoltà intrinseche associate alle ali idrodinamiche, gli scafi a propulsione umana non hanno la stessa potenzialità di velocità degli aeroplani a propulsione umana che hanno raggiunto velocità superiori ai 25 nodi.

Fino a poco tempo addietro tutti i record di velocità di scafi a propulsione umana erano detenuti da scafi a dislocamento con propulsione a remi. Nel 1984, nell'intento di superare queste velocità, due di noi (Brooks e Abbott) hanno progettato e costruito *Flying Fish I*, il primo scafo ad ali idrodinamiche capace di «volare» solo con propulsione umana. Lo spinoso problema di portare lo scafo al di là della velocità di decollo, che tormentava fin dall'inizio i progettisti, fu inizialmente superato eliminando la necessità di avere una carena dislocante. La velocità per volare fu raggiunta catapultando lo scafo nell'acqua da una piattaforma galleggiante grosso modo come vengono fatti decollare gli aerei dal ponte di una portaerei. Usando la partenza «lanciata», il ciclista Steve Hegg, medaglia d'oro olimpica, coprì, pedalando su *Flying Fish I*, una distanza di 2000 metri in 6 minuti e 38 secondi, migliorando di 11 secondi il record del mondo di un singolo rematore. Il tempo, naturalmente, non è direttamente paragonabile, perché il record di remata era ottenuto con partenza da fermo.

Flying Fish I ha un elica a pedali ad alto rendimento e due ali sottili sostenute da strette strutture verticali. L'ala principale, che porta il 90 per cento del peso dello scafo, ha una grande apertura alare (1,8 metri) per ridurre al minimo la resistenza indotta e corda, o larghezza, piccola per ridurre la resistenza di attrito. L'ala frontale più piccola ha una configurazione molto simile a una T rovesciata ed è soggetta a un carico leggero: il suo compito principale è quello di fornire stabilità e controllo. A questo scopo è collegata con un piccolo congegno a forma di spatola che controlla automaticamente la profondità dell'ala. Il congegno pattina sulla superficie dell'acqua regolando continuamente un sottile flap (simile al timone di profondità sulla coda di un aereo) al quale è collegato.

Il sostegno dell'ala frontale funge anche da timone ed è collegato al manubrio della bicicletta per potere virare. Lo scafo si guida allo stesso modo di una bicicletta. La struttura che normalmente sta al di fuori dell'acqua è infatti il telaio di una bicicletta modificato.

Flying Fish II è stato sviluppato migliorando la prima versione del nostro scafo. Abbiamo attaccato dei galleggianti leggeri sperando di poter decollare da fermo senza assistenza. Questo si è dimostrato possibile e, con l'allenamento, sono risultati necessari solo tre secondi per raggiungere la posizione di volo con partenza da fermo. Lo scafo è diventato così più pratico potendo anche «atterrare» e decollare sui suoi galleggianti. (Il *Flying Fish I*, che veniva catapultato, procurava al pilota un bel bagno quando questi smetteva di pedalare.)

A bordo del *Flying Fish II* uno di noi (Abbott) fece registrare il tempo record di 6 minuti e 39,44 secondi per coprire un percorso di 2000 metri con partenza da fermo, circa 10 secondi in meno del record di uno scafo con un solo rematore. Con partenza in velocità lo scafo ad ala idrodinamica ha permesso di coprire un percorso lanciato di 250 metri in 38,46 secondi, raggiungendo la velocità di circa 13 nodi.

Il tempo è maturo per una rivoluzione tecnologica negli scafi a propulsione umana. Laser International ha appena introdotto il Mallard, uno scafo navigante parzialmente coperto progettato da Garry Hoyt. Diversi nuovi catamarani a pedali e proa (scafi che hanno una carena principale e un piccolo bilanciere stabilizzante) offrono buona tenuta in condizioni di mare mosso e una velocità impressionante. Jon Knapp della Saber Craft ha progettato e costruito un proa con propulsione a elica che è più veloce di uno scafo monocarena in acque mosse, ma che, diversamente da quest'ultimo, non richiede particolare abilità per il governo. Il Dorycycle, un monoscafo a propulsione a elica progettato da Philip Thiel, offre buone capacità di carico a velocità doppie rispetto a quelle del dory a remi da cui è derivato.

A prescindere dal fatto che gli scafi ad ala idrodinamica possano o meno diventare popolari nel diporto, sembra fuori di dubbio che saranno alla ribalta nei prossimi miglioramenti di record. La International Human Powered Vehicle Association incoraggia la competizione fra i mezzi a propulsione umana su terra, in acqua e in aria, senza alcun limite arbitrario nella loro progettazione. Tale competizione spingerà sempre più in alto i valori della velocità degli scafi ad ala idrodinamica a propulsione umana. Non è esagerato prevedere scafi di questo genere in grado di raggiungere velocità addirittura dell'ordine dei 20 nodi: valori una volta e mezzo superiori alla velocità di *Flying Fish II*.

Un complesso di appartamenti dell'antica Roma

A Ostia, nelle case a giardino del II secolo d.C. dominava una geometria rigorosa: uno schema assai ricco di significati filosofici era alla base della progettazione sia della pianta in generale sia dei singoli mosaici

di Donald J. Watts e Carol Martin Watts

Vicino alla foce del Tevere, sotto il corridoio aereo dell'aeroporto Leonardo da Vinci, si trovano i resti in mattoni dell'antico porto romano di Ostia. Fra le rovine più affascinanti vi sono quelle di un complesso di appartamenti, empori e giardini, chiamato «case a giardino». Questo centro residenziale è impressionante anche perché rivela una complessa pianificazione e risulta sorprendentemente simile a sviluppi molto più vicini a noi. Persino allo stato attuale di ruderi - con muri non più alti di un piano e con erbacce che infestano spazi che un tempo dovevano essere splendidi giardini - trasmettono un senso di ordine e di pianificazione.

Nell'ambito di uno studio più ampio di antiche abitazioni romane, abbiamo analizzato nei particolari le case a giardino e abbiamo trovato la chiave per comprenderne il progetto: una singola geometria che si ripete a ogni scala, dalla configurazione generale degli edifici alla disposizione dei mosaici sui pavimenti. Lo schema, che si fonda sul quadrato e su un modo particolare di dividerlo, è chiamato «sezione sacra». La sezione sacra, assicurando rapporti proporzionali fra le parti del complesso e fra le parti e il tutto, conferisce al progetto unità e armonia. Per l'antico romano, e in particolare per l'ignoto architetto delle case a giardino di Ostia, esso potrebbe avere avuto un significato filosofico profondo.

Al suo apogeo, nel I e II secolo d. C., Ostia fu una città popolosa, con il fiume prospiciente percorso da un fitto traffico mercantile e i magazzini pieni di frumento, olio e vino destinati alla capitale dell'impero. Fu l'insabbiamento del fiume, insieme con il declino dell'Impero romano, a «uccidere» Ostia. Il delta del Tevere perse gradualmente la sua navigabilità, la costa avanzò verso occidente e la città rimase indietro. Nel IX secolo essa fu abbandonata e nel millennio successivo inondazioni periodiche ne seppellirono le rovine nel fango.

Il lento insabbiamento del Tevere creò a Ostia quel che la catastrofica eruzione del Vesuvio nel 79 d.C. aveva creato a Pompei e a Ercolano: una vera miniera di informazioni sull'antica architettura abitativa romana. (La continua attività di ricostruzione ha invece provocato la distruzione della maggior parte delle antiche abitazioni nella città imperiale). I vari siti riportati in luce sono complementari. A Pompei e a Ercolano esistono molti esempi ben conservati della tradizionale abitazione unifamiliare, la *domus*, con la tipica disposizione delle stanze attorno a un atrio centrale. Nella popolosa città di Ostia, invece, come nella stessa Roma, solo le persone molto ricche potevano permettersi una *domus*: le classi medie e inferiori vivevano in edifici ad appartamenti di tre-sei piani, chiamati *insulae*. Gli scavi condotti a Ostia nei primi decenni del secolo attuale hanno riportato in luce molte di queste *insulae*, di cui le case a giardino sono gli esempi più impressionanti.

Anche se non sappiamo chi le abbia progettate, è nota l'epoca in cui furono costruite: attorno al 128 d.C., durante il regno di Adriano. La costruzione di un nuovo porto artificiale nei pressi di Ostia da parte di Traiano, il predecessore di Adriano, aveva accresciuto l'importanza commerciale della città come porta d'accesso per l'importazione di merci a Roma. Per ospitare gli immigranti affluiti in gran numero nella città portuale (la cui popolazione aumentò a circa 50 000 unità), oltre metà di Ostia fu ricostruita; le *domus* furono sostituite da monumentali *insulae*, costruzioni a volta di calcestruzzo coperto da mattoni. Traiano e Adriano furono i patrocinatori di quest'attività edilizia, ma le case a giardino, il più grande progetto singolo, furono probabilmente il frutto di investimenti privati. Pare che fossero destinate a ricchi mercanti: erano situate vicino al mare e lontano dalle banchine fluviali piene di traffico; gli appartamenti erano grandi e ben decorati, e alcune unità erano lussuose.

Il complesso constava di un perimetro di edifici continuo, la cui forma irregolare era imposta con ogni probabilità da strade preesistenti. Esso racchiudeva un cortile rettangolare di circa 100 metri per 80 (*si veda l'illustrazione a pagina 91*), nel quale sorgevano due grandi edifici autonomi, ciascuno consistente di appartamenti standard, con quattro appartamenti al piano terra. Ogni appartamento aveva uno spazio centrale che fungeva anche da corridoio di accesso ai vari locali; esso prendeva il posto di un atrio e veniva chiamato *medianum*, cioè locale di mezzo. Ai due estremi di questo lungo locale luminoso, che dava sul muro esterno, vi erano altre due stanze principali. Varie camere da letto più piccole, senza finestre, erano adiacenti al *medianum*, da cui ricevevano luce e aria.

Gli edifici perimetrali comprendevano almeno altri nove appartamenti al piano terra oltre a una quarantina di empori di vendita. Tutti gli appartamenti al piano terra negli edifici in mezzo al cortile e la maggior parte di quelli compresi negli edifici periferici avevano una scala interna che conduceva a un mezzanino.

Le case a giardino furono riportate in luce a Ostia con il resto della città all'inizio del secolo. Il complesso di appartamenti, costruito attorno al 128 d.C., constava di un gruppo di edifici perimetrali che racchiudevano un cortile rettangolare in cui sorgevano due edifici. La fotografia dall'alto è stata eseguita sopra l'asse principale est-ovest del complesso; l'ingresso di facciata (rivolto a levante) è in basso. La «Casa delle Muse», nell'angolo a nord-est, è stata restaurata.

Rampe di scale separate che si aprivano direttamente sul cortile o sulla strada permettevano di accedere agli appartamenti situati ai piani superiori. Sappiamo poco di questi ultimi; solo il piano terra delle case a giardino è rimasto. Lo spessore dei muri indica che gli edifici avevano probabilmente quattro piani, compresi i mezzanini. A seconda delle dimensioni degli appartamenti dei piani superiori e dell'eventuale esistenza di mezzanini anche a tali piani, il complesso potrebbe avere avuto da 40 a 100 appartamenti. Esso dava alloggio probabilmente a un numero di persone compreso fra 400 e 700, inclusi i venditori che abitavano nei loro empori.

Il complesso è orientato approssimativamente secondo i quattro punti cardinali; la facciata e l'ingresso principale sono rivolti a levante. Data la forma irregolare del perimetro, era il grande cortile rettangolare a imporre un ordine geometrico al complesso. Il suo asse principale segue il corridoio compreso fra i due edifici nel cortile e collega l'ingresso principale a levante con un ingresso più piccolo a ponente. L'asse minore nord-sud è definito da passaggi coperti che attraversano entrambi gli edifici nel cortile; la sua minore importanza è indicata dal fatto che gli ingressi a nord e a sud sono spostati rispetto a esso. I due assi si intersecano al centro del complesso. È probabile che il centro fosse ulteriormente sottolineato dalla pianta generale (un punto centrale di convergenza dei vari elementi architettonici è un carattere distintivo della progettazione edilizia a Roma), ma non lo si può dire con sicurezza: tutto ciò che rimane dei giardini originari sono sei fontane, in prossimità dei lati orientale e occidentale del cortile, dalle quali i residenti attingevano l'acqua.

Tanto le fontane quanto gli edifici nel cortile risultano gli elementi chiave nell'ordinamento geometrico delle case a giardino. A mano a mano che misuravamo e ispezionavamo tutto il complesso, ci siamo resi sempre più conto della mano dell'architetto, di una volontà deliberata nella scelta delle dimensioni e della precisa posizione degli edifici nel cortile. Per spiegare quest'ordine, abbiamo preso in considerazione, per rifiutarle via via, decine di schemi geometrici sui quali l'architetto doveva essersi fondato e, alla fine, ci siamo imbattuti nell'idea della sezione sacra.

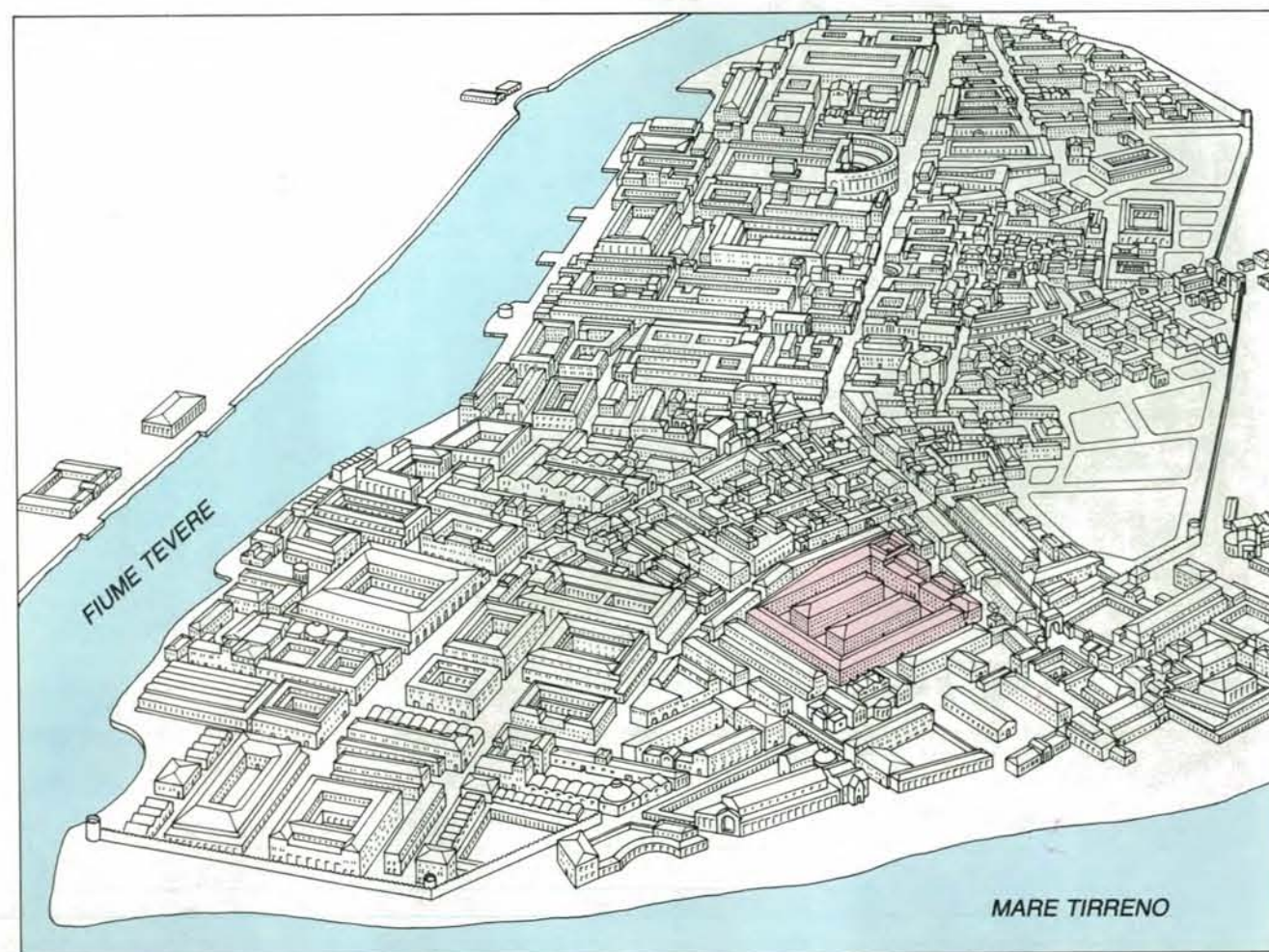
La sezione sacra è un'operazione geometrica semplice (*si veda l'illustrazione a pagina 92*), che può essere eseguita con riga e compasso, i quali, a quanto sappiamo dagli scritti di Vitruvio, erano usati entrambi dai costruttori romani nella definizione dei piani di base. Si comincia disegnando un quadrato (il quadrato di riferimento) e le sue diagonali. Poi, facendo centro negli angoli del quadrato, si tracciano quarte di cerchio, ciascuna con raggio uguale alla metà della diagonale. Questi quattro archi passano per il centro del quadrato e intersecano, ciascuno, due lati adiacenti, tagliando nel complesso ciascun lato in tre segmenti, dei quali il centrale è più grande degli altri. Congiungendo i punti di intersezione, il quadrato di riferimento è suddiviso in una griglia a nove parti, al centro della quale è un altro quadrato (il quadrato della sezione sacra), che funge da base per la successiva sezione sacra.

L'espressione sezione sacra non è antica; essa fu coniata una ventina di anni fa dallo studioso danese Tons Brunés. Si dà il caso che nella sezione sacra la lunghezza di un arco sia uguale, con un'approssimazione dello 0,6 per cento circa, alla lunghezza della diagonale del rettangolo che è pari alla metà del quadrato di riferimento. Secondo Brunés, questa quasi eguaglianza di un arco e di una linea retta potrebbe aver convinto gli antichi costruttori di aver trovato un modo per quadrare il cerchio, ossia per costruire un quadrato dal perimetro uguale alla circonferenza di un dato cerchio, e viceversa. (È impossibile quadrare esattamente un cerchio perché il perimetro di un quadrato è un numero razionale, mentre la circonferenza di un cerchio è proporzionale al numero irrazionale π.) Per gli antichi geometri il cerchio simboleggiava la parte spirituale, inconoscibile, dell'universo, e il quadrato rappresentava il mondo comprensibile. La quadratura del cerchio era un modo per esprimere l'inconoscibile attraverso il conoscibile, il sacro attraverso il familiare. Di qui il termine sezione sacra.

Benché Vitruvio non parli specificatamente di sezione sacra, menziona l'importanza degli schemi geometrici come modo per conseguire quella che considera la qualità più importante di un buon progetto: i rapporti proporzionali fra i vari elementi. Brunés ha ricostruito l'applicazione della sezione sacra dall'antico Egitto, attraverso i greci e i romani, all'Europa medievale. (Egli ritiene che l'idea sia stata trasmessa dall'Egitto in Grecia, nel VI secolo a.C., da Pitagora.) In particolare, Brunés ha trovato un'applicazione della sezione sacra a Roma, nel progetto del Pantheon, costruito pressappoco nello stesso periodo delle case a giardino di Ostia.

Le sezioni sacre sono onnipresenti nelle case a giardino, a cominciare dalla pianta generale del complesso. Una sequenza di tre sezioni di tre quadrati di riferimento definisce sia le dimensioni sia la posizione degli edifici del cortile e quindi anche degli spazi aperti nel cortile stesso (*si veda l'illustrazione a pagina 93*). Le sezioni sottolineano centro e assi del complesso di appartamenti, conferendo a quest'ultimo unità formale.

Il quadrato di riferimento più grande segue il perimetro del complesso, ma lo fa solo in modo approssimativo, perché il perimetro ha una forma irregolare. Che il quadrato non sia stato scelto in modo arbitrario si desume però dal fatto che un cerchio descritto al suo interno e tangente ai suoi lati passa esattamente per gli angoli del cortile. Il centro tanto del cerchio quanto del quadrato coincide anche con il centro del complesso. Le sezioni sacre dei lati orientale e occidentale del quadrato definiscono linee di guida che coincidono con i muri rivolti a nord e a sud dei due edifici nel cortile. I muri riaffermano il predominio dell'asse est-ovest stabilito dalla forma del cortile.

Procedendo verso il centro, il successivo quadrato di riferimento si fonda sulla larghezza del cortile. Poiché il cortile è rettangolare, i lati est e ovest del qua-

Ostia conobbe un enorme sviluppo nel II secolo d.C. La costruzione di un nuovo porto artificiale (non raffigurato) sul Tirreno aveva accresciuto l'importanza della città come porto principale di Roma; merci importate da vari paesi mediterranei venivano trasportate dal porto di Ostia a Roma, distante circa 25 chilometri, mediante navi che risalivano il Tevere. Quando la popolazione raggiunse le 50 000 unità, furono costruiti molti edifici ad appartamenti, fra i quali le case a giardino (*in colore*). Il disegno si basa su un modello derivato dalle rovine.

La pianta del piano terra delle case a giardino ne mette in luce la geometria basilare. L'asse principale est-ovest e l'asse secondario nord-sud si intersecano al centro del complesso. Gli edifici avevano probabilmente quattro piani. Ognuno dei 17 appartamenti del piano terra era formato da varie stanze che circondavano uno spazio centrale, probabilmente chiamato *medianum* (*in colore chiaro*). Il complesso comprendeva anche molti empori di vendita (*in colore scuro*). Gli ingressi ai giardini nel cortile si aprivano ai lati e agli angoli del complesso.

drato non seguono i lati del cortile; essi coincidono, invece, esattamente con il bordo interno delle sei fontane. Le sezioni sacre di questo quadrato definiscono gli stadi di costruzione successivi: gli assi longitudinali degli edifici nel cortile, lungo i quali spessi muri di confine separavano appartamenti contigui, affacciati in direzioni opposte. I muri di confine sottolineano ulteriormente l'orientazione est-ovest del complesso.

Il quadrato della sezione sacra, al centro del secondo quadrato di riferimento, svolge la funzione di terzo e ultimo quadrato di riferimento. Le sue sezioni sacre segnano a loro volta la posizione dei muri più interni degli edifici nel cortile, cioè dei muri più vicini al centro del complesso, che delimitano il corridoio centrale in direzione est-ovest. Quando si sovrappongono, i tre quadrati vengono a essere racchiusi concentricamente l'uno nell'altro e le loro sezioni sacre sembrano sfogliarsi come squame di cipolla.

La geometria delle sezioni sacre non solo unifica il complesso nella sua totalità, ma determina inoltre la pianta da singoli edifici. Per esempio, la diagonale del quadrato della sezione sacra più piccola nella sequenza menzionata sopra è uguale alla larghezza esterna degli edifici nel cortile. Inoltre, ciascun edificio è lungo cinque di questi quadrati. Il quadrato centrale nella fila dei cinque racchiude le rampe di scale e gli ingressi degli edifici, mentre i quadrati ai lati definiscono i locali di abitazione.

La geometria si estende ulteriormente ai singoli appartamenti degli edifici nel cortile (*si veda l'illustrazione a pagina 95*). In essi la sezione sacra serve non soltanto come metodo per determinare la posizione di taluni elementi del progetto, ma anche come fonte di una serie di dimensioni proporzionali, corrispondenti a numeri interi, che regolano la pianta degli appartamenti. In questo caso il quadrato di riferimento ha una diagonale di 58 piedi romani, che è la larghezza interna anziché quella esterna degli edifici del cortile, e un lato di 41 piedi romani. (Un piede romano è pari a 29,5 centimetri.) Due sezioni sacre successive forniscono segmenti di linea di 17, 12, sette e cinque piedi romani.

Queste dimensioni o loro multipli figurano con grande rilievo nella pianta degli appartamenti. La larghezza interna di un appartamento è pari a 28 piedi, ossia quattro volte sette piedi. La larghezza del *medianum* e delle camere da letto è metà di tale valore, ossia 14 piedi. Le finestre nella stanza più grande hanno la larghezza di cinque piedi e sono separate da uno spazio di due piedi, cosicché la larghezza di ogni unità comprendente una finestra è di sette piedi. I corridoi all'interno degli appartamenti sono larghi cinque piedi. La larghezza dello spazio comune, formato dal corridoio coperto e dalla rampa di scale che conduce ai piani superiori, è di 17 piedi.

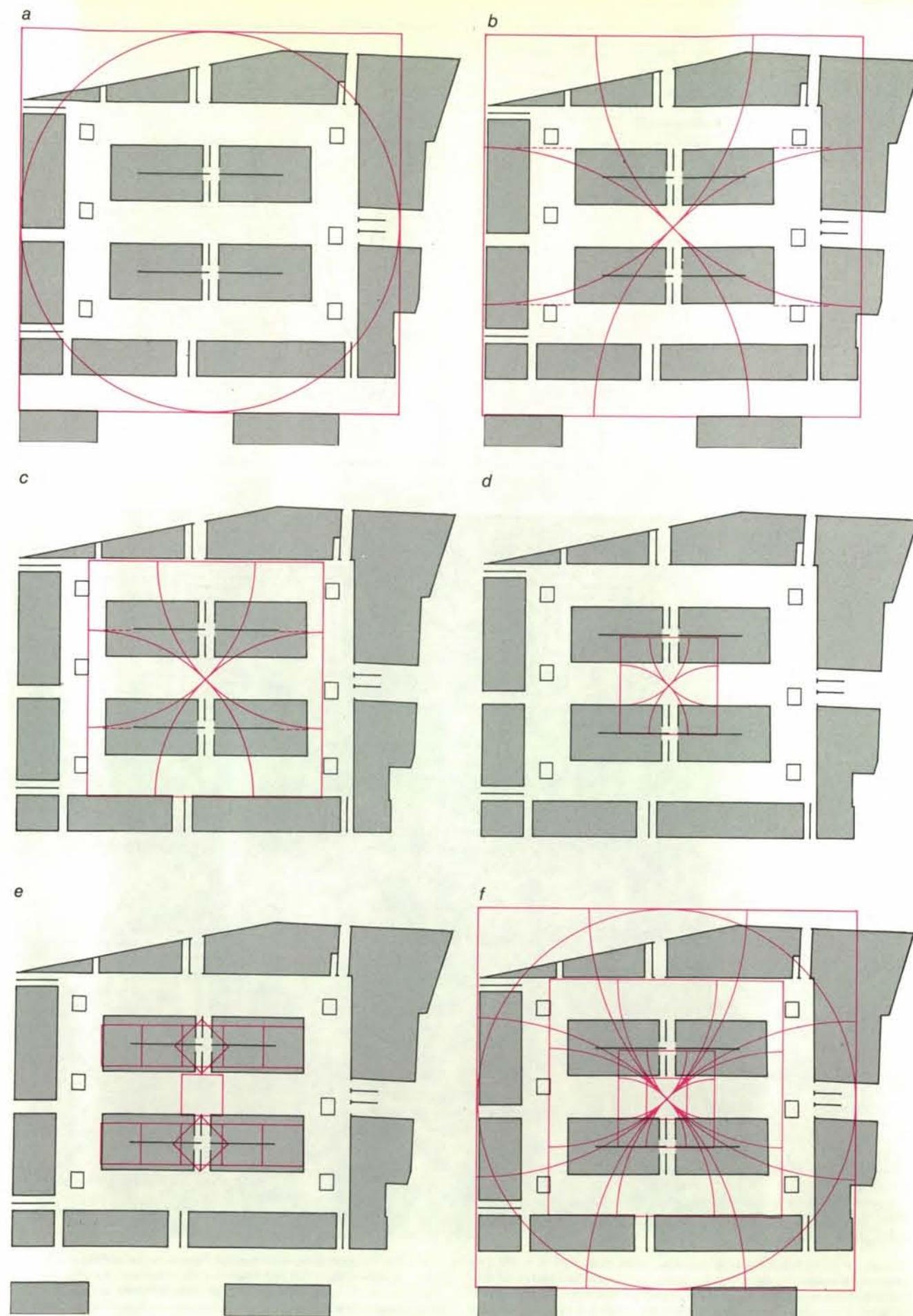
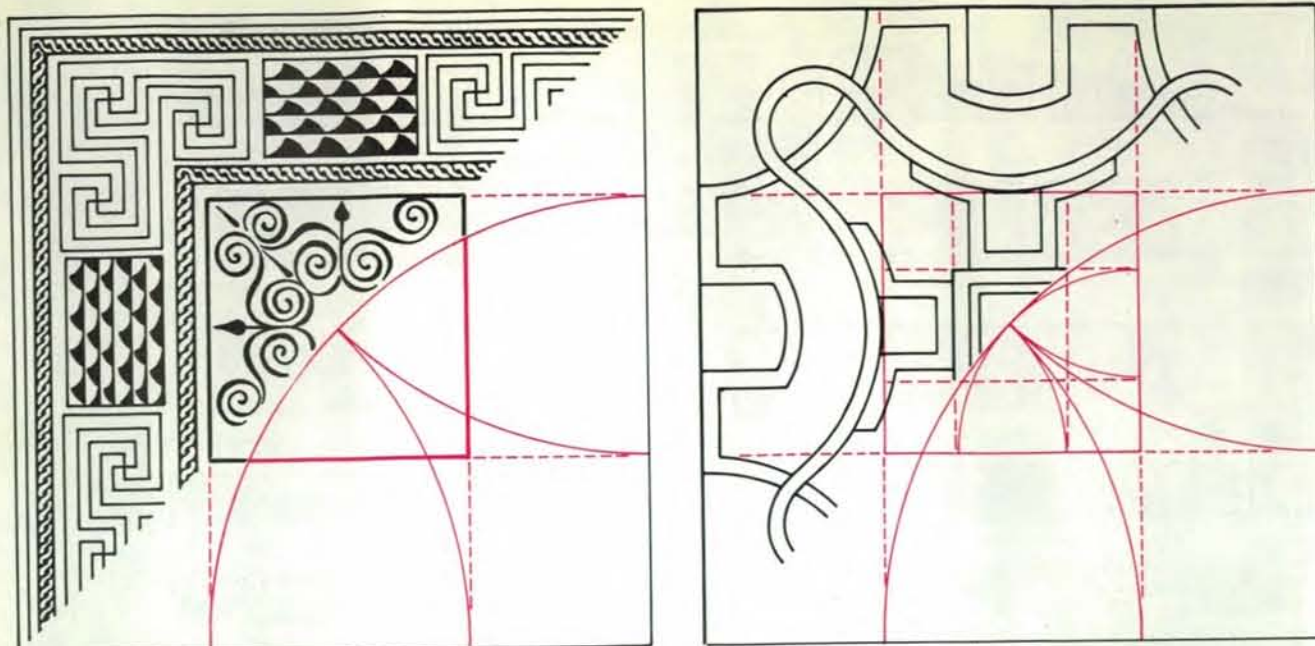
Le dimensioni devono avere avuto più che un significato semplicemente pratico per l'architetto delle case a giardino: esse sono i numeri generati dal procedimento pitagorico per approssimare la radice quadrata irrazionale di due. Il procedimento comincia con un quadrato di un'unità di lato, alla cui diagonale, che secondo il teorema di Pitagora è uguale alla radice quadrata di due, si assegna il valore approssimato di uno. La diagonale si aggiunge poi al lato del quadrato per generare il lato di un quadrato maggiore; la diagonale del secondo quadrato è pressappoco uguale alla somma della diagonale del primo quadrato più il doppio del suo lato. Costruendo così una serie di quadrati successivi e dividendo la diagonale approssimata di ogni quadrato per il lato dello stesso quadrato, si ottiene la serie di rapporti 1/1, 3/2, 7/5, 17/12, 41/29 e così via. La serie converge verso la radice quadrata di due.

La stessa serie di rapporti si può generare con quadrati e sezioni sacre, poiché anche il rapporto del lato di un quadrato al raggio dell'arco della sua sezione sacra è uguale alla radice quadrata di 2. (Il lato di ciascun quadrato nella serie si costruisce aggiungendo il lato del quadrato precedente al doppio della sua sezione sacra.) Il quinto quadrato nella serie, come il quadrato che è alla base della pianta degli appartamenti delle case a giardino, ha un lato di 41 e una sezione sacra di 29.

Fondando la pianta su una serie di numeri che si approssimano alla radice quadrata irrazionale di due, l'architetto delle case a giardino procedeva a un'asserzione filosofica affine a quella della quadratura del cerchio (operazione che equivale ad approssimare π). In entrambi i casi la sezione sacra è il mezzo per esprimere l'irrazionale e l'indefinibile con il razionale e il definibile.

Alla scala più piccola, nelle case a giardino, la sezione sacra è alla base del disegno delle singole decorazioni di pavimenti, pareti e soffitti. Nessuna delle decorazioni negli edifici del cortile si è ben conservata, mentre se ne sono conservate alcune negli edifici perimetrali. In un edificio detto Casa delle Muse, all'angolo nordorientale del complesso, un mosaico quadrato riempie la maggior

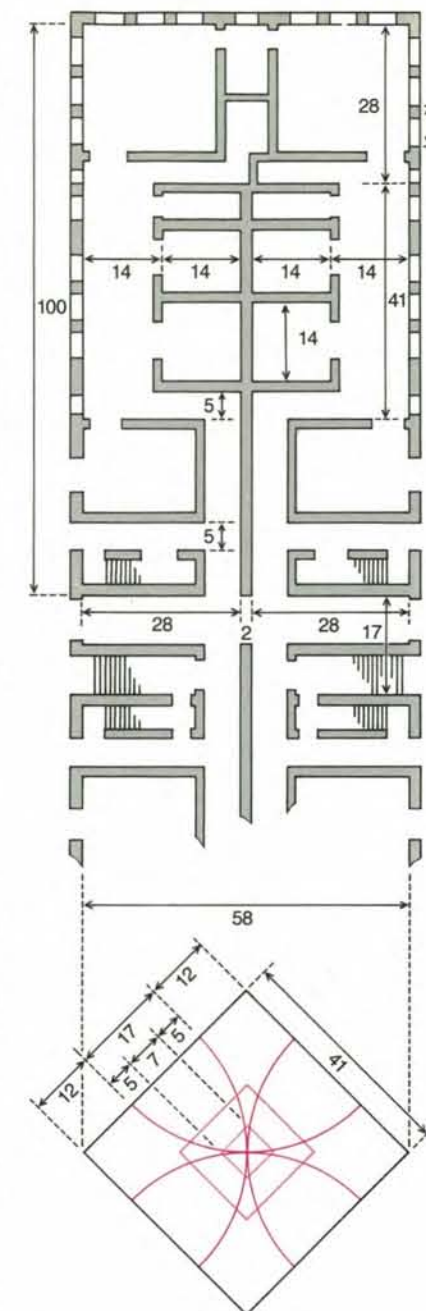
Le «sezioni sacre» di un quadrato di riferimento sono costruite tracciando archi che hanno il centro negli angoli del quadrato e passano per il suo centro (*in alto*). Unendo fra loro i punti nei quali gli archi intersecano i lati si ottiene una griglia divisa in nove parti: il quadrato centrale è detto quadrato della sezione sacra. La lunghezza di ogni arco *AB* è pari, con un'approssimazione minore dello 0,6 per cento, a quella della diagonale *CD* del rettangolo pari alla metà del quadrato di riferimento (*al centro*). La sezione sacra fornisce così un metodo approssimato per quadrare un cerchio: il perimetro di un quadrato composto da quattro segmenti *CD* è quasi uguale alla circonferenza di un cerchio composto dai quattro archi della sezione sacra (*in basso*).

L'ordine geometrico del complesso delle case a giardino di Ostia è definito da tre sezioni sacre successive. Un quadrato, che è grosso modo corrispondente con il perimetro del complesso, racchiude un cerchio che è tangente agli angoli del cortile (*a*). Le sezioni sacre dei lati orientale e occidentale di questo quadrato di riferimento determinano la posizione dei muri esterni degli edifici nel cortile (*b*). Il secondo quadrato di riferimento, concentrico al primo, è definito dalla larghezza del cortile e dalla posizione delle fontane; le sezioni sacre dei suoi lati orientale e occidentale guidano la posizione dei muri di separazione lungo gli assi longitudinali degli edifici nel cortile (*c*). Il terzo quadrato di riferimento è il quadrato della sezione sacra del secondo quadrato e le sue sezioni definiscono i muri più interni degli edifici nel cortile (*d*). Gli edifici sono esattamente cinque volte più lunghi dell'ultimo quadrato della sezione sacra e la loro larghezza è pari alla sua diagonale (*e*). Una sovrapposizione di tutte le sezioni sacre mostra come esse abbiano un centro comune, sottolineando in tal modo l'asse principale est-ovest del complesso (*f*).

Mosaici e pitture nelle case a giardino sono realizzati in molti casi secondo la geometria della sezione sacra. La fotografia illustra un mosaico pavimentale nella Casa delle Muse. Il medaglione quadrato al centro è il quadrato della sezione sacra del bordo esterno (*a sinistra in alto*). Nel disegno della decorazione dipinta di un soffitto (*a destra in alto*), le sezioni sacre del bordo esterno determinano il raggio degli archi d'angolo, e i grandi archi laterali sono tangenti al quadrato della sezione sacra. Le sezioni di questo incorniciano il medaglione centrale.

parte del pavimento di una stanza. Al centro vi è un medaglione quadrato, le cui dimensioni sono definite dalla sezione sacra del quadrato esterno del mosaico. Composizioni simili si trovano in altre parti delle case a giardino. In alcuni casi i mosaici in stanze diverse dello stesso appartamento sembrano connessi per dimensioni e geometria. Per esempio, il mosaico in una stanza può avere le dimensioni del quadrato della sezione sacra del mosaico di una stanza più grande.

Oltre ad avere mosaici sui pavimenti, è probabile che la maggior parte delle stanze nelle case a giardino fosse decorata con pitture dai colori vivaci sul soffitto e sulle pareti. Nessuna delle pitture dei soffitti si è conservata intatta, ma di una possediamo un numero di frammenti abbastanza grande da permetterci di ricostruirne la composizione. Come il mosaico sul pavimento, essa è quadrata e organizzata attorno a un centro. Sezioni sacre del quadrato esterno determinano la posizione di due insiemi di archi; il quadrato della sezione sacra del perimetro esterno è poi tagliato di nuovo per definire un medaglione centrale.

Anche le pitture sulle pareti interne ci forniscono attestazioni della sezione sacra. Certo, molte pareti non sono quadrate, essendo o più lunghe che alte o più alte che lunghe. Nondimeno, in quasi tutte le pitture parietali conservate, la composizione sembra fondarsi su un quadrato di riferimento concettuale il cui lato corrisponde alla lunghezza della parete. La geometria della sezione sacra non solo determina la posizione degli elementi principali della composizione, ma fornisce inoltre, come nella pianta degli appartamenti, una serie di numeri che governano le dimensioni nel dipinto, aventi importanza cruciale.

La sezione sacra è solo uno dei molti schemi geometrici, perloppiù fortemente centralizzati e assiali, che si trovano nelle antiche progettazioni romane. Eppure la sua ripetizione insistente a tutte le scale nelle case a giardino di Ostia induce a pensare che avesse una sorta di importanza ritualistica. Stiamo ora cercando la sezione sacra anche altrove nell'architettura romana. Le nostre misurazioni ne hanno già rivelato la presenza nelle piante e nelle decorazioni di molte case unifamiliari a Pompei e a Ercolano, costruite da uno a tre secoli prima delle case a giardino; a quanto pare la sezione sacra era una caratteristica della *domus*, ancora viva quando gli architetti romani cominciarono a progettare edifici ad alta densità abitativa. È ovviamente l'applicazione della sezione sacra andò oltre la progettazione di abitazioni. Come già ricordato, Brunés la scoprì in edifici pubblici come il Pantheon. Essa potrebbe essere stata usata persino nella pianificazione urbanistica: abbiamo trovato prove che un'intera città romana, costruita in Medio Oriente nel I secolo d.C., fu progettata secondo una geometria basata sulla sezione sacra.

X	Y	Y/X ($\sqrt{2}$=1,14142...)
1	1	1
2	3	1,5
5	7	1,4
12	17	1,4166
29	41	1,4137

Le piante degli appartamenti negli edifici del cortile sono regolate dalle sezioni sacre di un quadrato i cui lati sono lunghi 41 piedi romani (un piede romano è pari a 29,5 centimetri) e la cui diagonale è pari alla larghezza interna degli edifici (58 piedi). I due appartamenti contigui qui raffigurati sono separati da un muro di due piedi. In essi si notano le dimensioni determinate mediante successive sezioni sacre del quadrato: sono particolarmente frequenti il numero 7 e i suoi multipli. La serie di numeri generati dalle sezioni sacre è la stessa usata dagli antichi geometri per approssimare la radice quadrata irrazionale di due.

ARCHITETTURA E ARTE

LE SCIENZE *edizione italiana di* **SCIENTIFIC AMERICAN**

ha dedicato all'argomento diversi articoli:

L'origine dell'ambiguità nelle opere di Maurits C. Escher
di M. L. Teuber (n. 75)

Pieter Bruegel il Vecchio e la tecnica del Cinquecento
di H. A. Klein (n. 117)

La conservazione della pietra
di K. L. Gauri (n. 120)

Norme architettoniche nella Cina del XII secolo
di E. Glahn (n. 155)

L'architettura di Christopher Wren
di H. Dorn e R. Marck (n. 157)

Intarsi rinascimentali: l'arte della geometria
di A. Tormey e Y. Farr Tormey (n. 169)

Le volte a ventaglio
di W.C. Leedy, Jr. (n. 176)

Le chiese di legno della Norvegia
di P. Anne, R.L. Sack e A. Selberg (n. 182)

Il Crystal Palace
di F. T. Kihlstedt (n. 196)

Sperimentazione strutturale nell'architettura gotica
di R. Mark e W. W. Clark (n. 197)

Il restauro delle vetrate medioevali
di G. Frenzel (n. 203)

(RI)CREAZIONI AL CALCOLATORE

di A. K. Dewdney

Montagne frattali, piante graftali e altra grafica al calcolatore della Pixar

Posso facilmente immaginare il primo vero lungometraggio generato dal calcolatore. È l'anno 1991. Inciampo camminando nel corridoio tra le poltrone con in mano uno smisurato cartoccio di popcorn sintetico e una bibita contenente alcuni additivi che rendono inutili i normali ingredienti. Le luci si abbassano, il sipario si apre e lo schermo argenteo si anima con una riduzione della trilogia di J. R. R. Tolkien *Il signore degli anelli*. Lo hobbit Frodo passeggia per una stretta valle. In lontananza, i picchi innevati di una montagna si alzano frastagliati verso il cielo. In primo piano, alberi esotici e piante di specie sconosciute brillano alla luce del sole. La scena cambia e compare un mago che scruta in una sfera di cristallo. Al centro della sfera si vede una fortezza con gli spalti merlati avvolti dalle fiamme.

È difficile dire con esattezza quanto potrà essere convincente in questo film Frodo che cammina e parla, ma sono sicuro che le montagne, le piante, la sfera di cristallo e le fiamme riusciranno tutte magnificamente. Il successo sarà dovuto ampiamente al software e all'hardware avveniristici di una società di nome Pixar, già Lucasfilm Computer Graphics Laboratory. Dopo aver visitato questo affascinante centro della grafica al calcolatore di San Rafael in California, posso mettere a parte i lettori dei più riposti segreti di montagne e alberi. Chiunque possieda un calcolatore domestico è ora in grado di generare immagini che assomigliano molto a questi oggetti. La limitazione di spazio che la rubrica impone mi impedisce di trattare estesamente della sfera di cristallo e del fuoco; svelerò, però, i principi di base per generarli.

Nell'ipotetico film descritto, la telecamera potrebbe zoomare sulle cime innevate alle spalle di Frodo. Non si potrebbe vedere un ammasso più spaventoso: ogni vetta è formata da vette più piccole e così via: un regresso all'infinito di piccoli picchi. Perfino un Orco, quella bestia mostruosa dai piedi di cuoio, non si troverebbe a proprio agio su quegli scabri pendii.

In linea di principio è facile generare una catena montuosa di questo aspetto. Per semplificare, ammettiamo che il terreno copra un'area triangolare. Si suddivide allora il triangolo in quattro triangoli più piccoli, trovando il punto di mezzo di ciascun lato e congiungendo i nuovi punti con tre segmenti. Ciascun triangolo viene a sua volta suddiviso nella stessa maniera. Si continua il procedimento fino a raggiungere i limiti di risoluzione o del tempo di calcolo. Il risultato - un reticolo di triangoli piuttosto monotono - può essere ravvivato aggiungendovi un po' di movimento in verticale: ogni volta che si aggiunge alla scena un nuovo punto medio, lo si sposta verso l'alto o verso il basso di una misura casuale. Gli spostamenti casuali, che in generale devono essere ridotti a mano a mano che i triangoli diventano più piccoli, trasformano i triangoli in vette frastagliate che si alternano a valli (*si veda l'illustrazione in alto di pagina 100*).

Perché questa tecnica dovrebbe produrre montagne dall'aspetto naturale? La risposta può risiedere in parte nel fatto che il procedimento produce un frattale: un tipo di oggetto che crescendo rivela un maggior numero di dettagli. A quanto pare, in tutta la natura si possono vedere frattali. Benoit B. Mandelbrot, l'infaticabile studioso di frattali del Thomas J. Watson Research Center della IBM a Yorktown Heights, New York, utilizza linee costiere per illustrare l'idea di fondo. Immaginiamo che ci sia chiesto di misurare la costa francese con un'asta lunga un chilometro. Facendo ruotare l'asta sulle sue estremità, con una faticosa marcia lungo la costa si arriva a calcolare il numero di chilometri. Molte piccole baie e molti piccoli promontori, però, vengono tralasciati e la lunghezza finale misurata in questo modo non è del tutto esatta. Se si ripete l'esercizio con un righello da un metro si ottiene una misura più precisa e più lunga. Anche in questo caso, però, viene trascurato un gran numero di minuscole insenature e lingue di terra. Senza dubbio, un righello da un centimetro sarebbe più preciso.

Come regola generale possiamo dire che la lunghezza della costa misurata aumenta con il ridursi dell'asta di misurazione. La relazione tra la lunghezza misurata e la dimensione dell'asta è un particolare numero detto «dimensione frattale». A differenza di una comune dimensione, una dimensione frattale di solito è espressa sotto forma di frazione, non di numero intero. La linea costiera in questione potrebbe avere, per esempio, dimensione frattale pari a 3/2. Questa forma può essere vista come una via di mezzo tra una forma a una dimensione (una linea retta) e una forma a due dimensioni (un piano). Se una linea costiera fosse relativamente diritta, la sua dimensione frattale sarebbe vicina a 1; se invece fosse molto frastagliata, la sua dimensione frattale si avvicinerebbe a 2 come se cercasse di riempire un piano a due dimensioni.

Il modello frattale della natura implica un regresso infinito di dettagli. Dal punto di vista della grafica al calcolatore, la questione del regresso all'infinito non si pone; è sufficiente che il paesaggio appaia dettagliato a tutti i livelli di ingrandimento. Fino ai limiti di risoluzione dello schermo, le montagne da generare hanno caratteristiche con finezza pari ai triangoli finali usati nella suddivisione descritta in precedenza. Anche se l'algoritmo completo per disegnare montagne è troppo lungo e complesso per poterlo descrivere in questa sede, c'è un semplice programma chiamato MOUNTAIN che disegna il monte Mandelbrot in sezione trasversale. MOUNTAIN illustra l'idea fondamentale di punti di suddivisione a spostamento casuale lungo un asse verticale. L'artista frattale inizia con un unico segmento orizzontale. Si determina il punto di mezzo e lo si sposta su o giù di una misura casuale. Ciascuno dei due segmenti risultanti viene poi suddiviso e perturbato. Il procedimento può essere proseguito in maniera analoga alla tecnica di suddivisione dei triangoli.

MOUNTAIN conserva due matrici, dette *punti* e *linee*, per seguire il profilo montuoso. Ciascuna matrice ha due colonne e un numero di righe adeguato alla risoluzione dello schermo (per esempio 2048). Le due colonne di *punti* contengono coordinate e le due colonne di *linee* contengono indici; ciascuna linea è specificata come coppia di posizioni nella matrice *punti* che designa le coordinate delle estremità della linea. Dato che è interessante osservare come le suddivisioni che si susseguono formino il profilo di una montagna a partire da un poligono apparentemente poco promettente, MOUNTAIN mette ogni generazione sotto il controllo dell'utente. Al termine di un singolo ciclo principale, il programma chiede all'utente se vuole un'altra iterazione. Se la risposta è positiva, l'esecuzione torna all'inizio del programma.

Il ciclo principale trasforma gli insiemi attuali di punti e linee in nuovi insiemi grandi il doppio. Per ottenere questo risultato scorre *linee* una riga alla volta, cerca gli indici dei punti corrispondenti e richiama le loro coordinate da *punti*. Con le coordinate delle estremità di una data linea, il programma calcola le coordinate del punto di mezzo della linea stessa, modificando casualmente, nel corso del procedimento, la coordinata y. L'algoritmo che segue fornisce una base adeguata per un programma. Le variabili j e k indicano le righe di *punti* e *linee* che a un dato momento vengono riempite con gli ultimi risultati della suddivisione. Le variabili *pt* e *ln* registrano il numero di punti e il numero di linee che formano la montagna prima che il programma entri nel ciclo principale. All'inizio j è uguale a *pt* e k è uguale a *ln*. L'indice i va da 1 a *ln*.

```
j ← j + 1
k ← k + 1
a ← linee (i,1)
b ← linee (i,2)
x1← punti (a,1)
y1← punti (a,2)
x2← punti (b,1)
y2← punti (b,2)
punti (j,1) ← (x1 + x2)/2
punti (j,2) ← (y1 + y2)/2 +
                    a caso (intervallo)
linee (i,2) ← j
linee (k,1) ← j
linee (k,2) ← b
```

Questa parte di MOUNTAIN si spiega in gran parte da sé. Una volta calcolate le coordinate del j-esimo punto, l'indice j viene memorizzato come secondo punto dell'i-esima linea e primo punto della k-esima linea. Il primo punto della linea i-esima è lo stesso di prima e il secondo punto della linea k-esima è identico al secondo punto originario della linea i-esima, vale a dire quello con indice b.

Al termine del ciclo di calcolo, *pt* e *ln* devono essere riportati, rispettivamente, agli ultimi valori di j e di k. Alla variabile *intervallo* inizialmente l'utente dà come valore la massima quantità di casualità verticale che può essere data al punto di suddivisione. Ogni volta che il ciclo viene completato, questa variabile deve essere divisa per 2 in modo che le fluttuazioni casuali siano sempre in scala con la dimensione delle caratteristiche da variare. La funzione *a caso* (*intervallo*) intende esprimere la selezione di un numero casuale compreso tra 0 e il valore di *intervallo* (al momento dato).

Se appaiono efficaci le montagne che stanno alle spalle di Frodo, gli alberi e le piante che lo circondano non sono da meno. Sono allo stesso tempo realistici e fiabeschi. Sembrano reali perché hanno ramificazioni simili a quelle delle vere piante, e fiabeschi perché non appartengono a specie familiari; il progettista grafico ha a disposizione un tale numero di parametri da non poter resistere alla tentazione di creare qualcosa di nuovo.

Le nuove «specie» sono denominate piante graftali, perché sono basate su grafi e hanno un'implicita natura frattale. Per implicita natura frattale intendo il fatto che le regole per generare la topologia di base delle piante potrebbero essere (ma non lo sono) applicate al limite di risoluzione dello schermo. In breve, un ramo non si sviluppa in un regresso all'infinito di rametti. Una volta sviluppato, il grafo che costituisce la base di una pianta può essere trasformato in una miriade di specie convincenti se lo si interpreta in termini di dimensione, colore, spessore, struttura e così via.

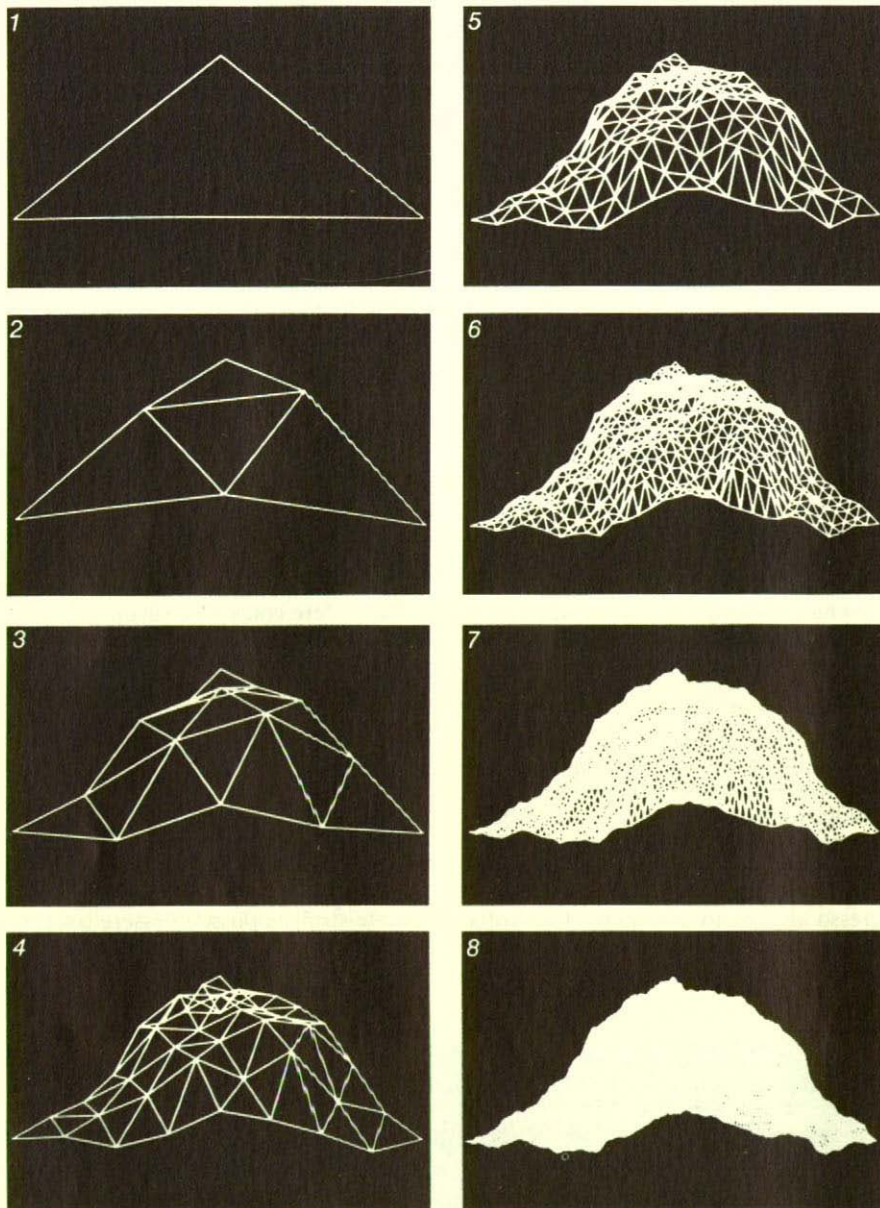
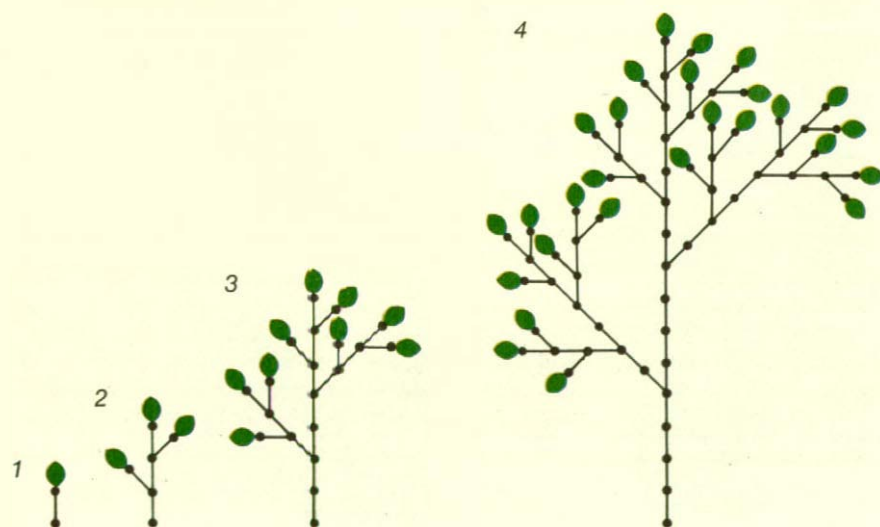
I grafi che sottendono una data pianta sono prodotti da sistemi L, una classe di grammatiche introdotta nel 1968 dal biologo e matematico danese Aristid Lindenmeyer. Un sistema L è in sostanza un insieme di regole per derivare nuove stringhe di simboli da vecchie stringhe. Le regole comprendono sequenze di sostituzione di simboli per singoli simboli. Per esempio, usando i numeri 0 e 1 e i simboli di parentesi quadra aperta e chiusa si può generare una vasta gamma di forme botaniche complesse con le seguenti regole:

```
0 → 1 [0]1[0]0
i → 1 1
[ → [
] → ]
```

Per vedere come funzionano le regole, supponiamo di partire con la stringa formata da un unico 0. A ogni simbolo di sinistra della stringa si sostituisce il suo corrispondente simbolo di destra in modo da ottenere le seguenti stringhe in successione:

```
0
1[0]1[0]0
1 1[1[0]1[0]0]1 1[1[0]1[0]0]1[0]1[0]0
```

Queste stringhe possono essere trasformate in grafi a forma di albero trattando ogni numero (0 o 1) come un segmento

Montagne frattali generate dal calcolatore imitano la natura

La suddivisione di triangoli produce una montagna

Quattro generazioni di una pianta graftale

e ogni parentesi come un punto di ramificazione. I segmenti 0 e i segmenti 1 sono di uguale lunghezza; tipicamente si distinguono perché si lasciano spogli tutti i segmenti 1, mentre si pone una foglia all'estremità esterna di ogni segmento 0.

Per esempio, lo stelo (o tronco) della stringa 1[0]1[0]0 è formato dai tre simboli che non si trovano tra parentesi; un segmento 1 è sormontato da un secondo segmento 1 e poi ancora da un segmento 0. Due rami, ciascuno formato da un singolo segmento 0, germinano da questa formula. Il primo ramo ha l'attaccatura sopra il primo segmento e il secondo sopra il secondo segmento. Prima di studiare l'illustrazione in basso a sinistra, i lettori potrebbero divertirsi a disegnare un po' di generazioni della struttura. Per amore di realismo, si possono aggiungere al modello altre caratteristiche interpretative. Si potrebbe specificare che per qualunque tronco dato (indipendentemente dal fatto che sia o meno il tronco principale) i rami dovrebbero dipartirsi alternativamente verso destra e verso sinistra. Non volendo imporre alcunché di più arduo ai lettori che desiderassero programmare graftali, sono lieto di suggerire, per le piante, semplici ramoscelli. I professionisti della Pixar trasformano la grammatica appena descritta in bellissime piante come quelle dell'illustrazione in alto della pagina a fronte.

Un programma in due parti detto PLANT genera l'n-esima stringa della successione precedentemente illustrata e poi la presenta come disegno al tratto. Nella sua prima fase, PLANT conserva le stringhe che genera in due matrici di simboli dette *stringaA* e *stringaB*. Ciascuna generazione di piante occupa una delle due matrici in modo alternato: la generazione di una matrice è derivata dalla generazione precedente dell'altra. Non è strettamente necessario immagazzinare simboli in queste matrici: se il programma esegue correttamente le sostituzioni, andranno benissimo anche i numeri 0, 1, 2 e 3.

Le regole del sistema L sono inglobate in enunciati condizionali. Per esempio, si può adattare il seguente passo di codice algoritmico per trasformare uno 0 nella i-esima posizione di *stringaA* in nove nuovi simboli di *stringaB*:

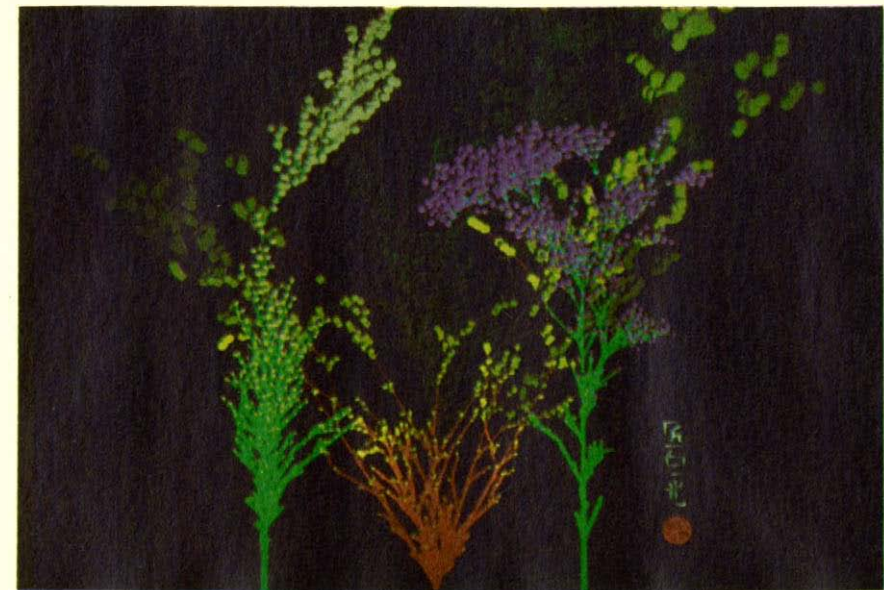
```
se stringaA(i) = 0, allora
    stringaB(j) ← 1
    stringaB(j + 1) ← 2
    stringaB(j + 2) ← 0
    stringaB(j + 3) ← 3
    stringaB(j + 4) ← 1
    stringaB(j + 5) ← 2
    stringaB(j + 6) ← 0
    stringaB(j + 7) ← 3
    stringaB(j + 8) ← 0
    j ← j + 9
```

Qui 0 e 1 stanno per se stessi, mentre 2 e 3 stanno, rispettivamente, per [e]. Se l'i-esimo simbolo di *stringaA* è 0, allora il programma inserisce la successione 1, 2, 0, 3, 1, 2, 0, 3, 0 in nove posizioni successive della matrice *stringaB* a partire dall'indice j (la prima posizione della seconda matrice che non è stata ancora riempita). Un unico ciclo nella prima fase di PLANT contiene quattro enunciati condizionali di questo genere, uno per ogni possibile simbolo incontrato. Il ciclo usa l'indice j come riferimento al simbolo della generazione che si sta elaborando. Il ciclo viene eseguito per il numero di generazioni voluto dall'utente. A ogni stadio, PLANT può chiedere all'utente se vuole un'altra (più lunga) stringa di simboli.

La seconda fase di PLANT, quella grafica, trasforma in un disegno la stringa prodotta dalla prima fase. L'operazione è compiuta ricorsivamente. Finché non incontra una parentesi sinistra, o 2, disegna una successione di segmenti in una data direzione. Quando viene presa in esame una parentesi sinistra di una data coppia, il programma disegna il successivo segmento in una nuova direzione, spostata di 45 gradi in senso antiorario rispetto a quella precedente. La fine del procedimento è segnalata dalla comparsa della corrispondente parentesi destra; qui può essere disegnata una foglia (di forma e colore completamente affidati alla fantasia del lettore). La comparsa di una seconda parentesi sinistra provoca la ripetizione del procedimento, solo che ora la nuova direzione è di 45 gradi in senso orario. Il resto è automatico.

PLANT utilizza un fattore di scala che dipende dalla complessità della pianta da disegnare. L'n-esima generazione, per esempio, è alta approssimativamente 2^n segmenti. Se lo schermo ha un'altezza di 200 pixel, i segmenti devono essere più corti di $200/2^n$. Senza dubbio, i lettori ambiziosi cercheranno varianti nella grammatica generativa, negli angoli di ramificazione e nella forma delle foglie. Se si eseguono queste varianti sullo stesso schermo, appariranno paesaggi di piante e alberi (non molto realistici, bisogna ammettere).

La sfera di cristallo dell'ipotetico film tratto da Tolkien sarebbe realizzata con una tecnica chiamata tracciamento di raggi (*ray tracing*); i merli in fiamme sarebbero simulati seguendo il movimento di un grosso sistema di particelle.

Il tracciamento di raggi richiede di specificare sia la geometria tridimensionale di una scena sia la posizione di una sorgente di luce. Quando lascia una sorgente, la luce si imbarca in una complicata storia di riflessioni e rifrazioni. L'occhio di un osservatore che si trovi sulla scena intercetterà alcuni raggi di luce che fluiscono dalla sorgente ma ne mancherà molti altri, in realtà la maggior parte. Per non sprecare tempo e potenza di calcolo, la tecnica del tracciamento di raggi lavora nella direzione opposta. Immaginiamo per un momento che la luce lasci invece l'occhio. Un ampio fascio di raggi si distende a ventaglio nella scena. Se un raggio colpisce una superficie riflettente o rifrangente, saetta via in una nuova direzione determinata dalle leggi dell'ottica. Infine il raggio colpisce una superficie assorbente, assumendo il colore ivi assegnato. Quel colore è registrato nel pixel corrispondente alla direzione del raggio di partenza.

Piante graftali prodotte alla Pixar

Un'immagine di palle da biliardo generata dal calcolatore illustra la tecnica di tracciamento di raggi

Qui sopra si può vedere un'immagine generata con questa tecnica. I raggi tracciati nella scena composta da palle di biliardo hanno una storia relativamente semplice. Nonostante la semplicità, però, possiamo vedere riflessi l'interno di una sala da gioco e un uomo che, in piedi con una stecca in mano, osserva il colpo.

Il grande sistema di particelle che si potrebbe usare per generare i merli in fiamme è lo sviluppo logico dei piccoli ammassi di punti che rendono simbolicamente le esplosioni in miniatura nei videogiochi. Alla Pixar, però, un sistema di particelle è molto più raffinato. All'interno di una certa regione, moltissime particelle vivono, si muovono e conducono la loro esistenza. Sotto il controllo del calcolatore, ciascuna particella è un punto che si sposta secondo una dinamica predeterminata. Nata in un certo istante, può muoversi per un po', magari dando anche vita a nuove particelle; poi può morire.

Sistemi di particelle sono stati usati in modo spettacolare in una scena del film *Star Trek II*. Una bomba Genesi viene sganciata su un pianeta morto pieno di crateri. La bomba crea un anello di stra-

Una scena dalla sequenza Genesi in **Star Trek II**

ne fiamme scintillanti che alla fine inghiottono il pianeta. Quando infine si spengono, vediamo la superficie trasformata in una lussureggiante biosfera. L'effetto è stato prodotto dalla Pixar nella sua precedente incarnazione come Lucasfilm Computer Graphics Laboratory. L'anello di fuoco in espansione era formato da sistemi di particelle in cui alcune particelle davano origine a interi nuovi sistemi. I nuovi sistemi rappresentavano particelle che, scagliate verso l'alto dalla superficie del pianeta, cambiavano colore e persino ricadevano sotto l'influenza della gravità.

Alvy Ray Smith, che dirige il settore ricerca e sviluppo alla Pixar, mi ha accompagnato in un giro dell'azienda durante la mia visita a San Rafael. Smith è già noto ai lettori di «Le Scienze» per il suo lavoro con gli automi cellulari (si vedano le rubriche «Giochi matematici» del maggio 1971 e «(Ri)creazioni al calcolatore» dell'ottobre 1985). Oltre che con Smith, pioniere nell'applicazione del metodo frattale alla vita delle piante col calcolatore, ho avuto un incontro con Loren Carpenter, specialista di montagne frattali, con Robert L. Cook, esperto del tracciamento di raggi, e con William Reeves, creatore dei sistemi di particelle. Nel bel mezzo di una discussione sul software grafico, Smith mi ha sorpreso affermando che l'attività principale dell'azienda non è tanto la produzione di effetti speciali per Hollywood quanto la costruzione di un calcolatore dedicato alla grafica e chiamato, abbastanza naturalmente, Pixar Image Computer.

Al cuore del Pixar Image Computer c'è una memoria da 24 megabyte per 2000 × 2000 pixel. È una risoluzione più che sufficiente per la maggior parte delle applicazioni. Ciascun pixel, per di più, è rappresentato da 48 bit di memoria, sufficienti a conservare copiose informazioni sul colore e la trasparenza. La grande memoria del Pixar è controllata da quattro elaboratori paralleli ad alta velocità, totalmente programmabili, che possono eseguire circa 40 milioni di istruzioni al secondo, una velocità che è di molti ordini di grandezza superiore a quella dei comuni calcolatori. Un'unità video comunica con la memoria a una velocità di 480 milioni di byte al secondo.

I primi Pixar sono stati messi in commercio nel maggio scorso. Sono destinati all'elaborazione di immagini in campo medico, al rilevamento a distanza, al disegno tecnico e all'animazione. Forse saranno usati anche per generare il mio ipotetico film.

La tappezzeria per la mente, argomento dell'articolo del novembre scorso, riguardava immagini al calcolatore quasi, ma non del tutto, ripetitive: un tipo di tappezzeria che ancora non si è visto. In effetti, però, configurazioni simili a queste erano conosciute una generazione fa. Michael Rossman, scrittore e commentatore politico che vive a Berkeley, California, coniò la frase «tappezzeria per la mente» nel 1971. Si riferiva ai delicati disegni che si vedono a occhi chiusi dopo l'assunzione di LSD. Scrive a questo proposito Rossman: «"Sembrava" una tappezzeria, mai a strisce ma con motivi ripetuti: di solito costellazioni di pappagalli, stelle marine, lampadine, fiocchi di neve, innominabili macchie di Rorschach, cangianti con sfrenata creatività... le immagini si trasformavano a catena una nell'altra, come se qualche stampa lineare infinita di Escher con immagini di metamorfosi si fosse animata in un film.»

I lettori ricorderanno che il più semplice tra i programmi per tappezzeria descritti in questa rubrica era stato realizzato da John E. Connett dell'Università del Minnesota. Secondo le sue indicazioni, le coordinate di ciascun pixel vengono elevate al quadrato e sommate. Il risultato è troncato a un intero. Se l'intero è pari, il pixel viene colorato di nero; se è dispari il pixel viene colorato di bianco. La comparsa di una tappezzeria (ripetizioni orizzontali e verticali) può essere spiegata in parte come fenomeno moiré: nella precedente descrizione sono implicite due configurazioni e la loro sovrapposizione crea l'effetto. La prima configurazione è la griglia rettangolare di pixel che costituisce lo schermo di visualizzazione. La seconda configurazione è una serie di anelli concentrici che rappresentano i punti del piano per i quali il procedimento di Connett genera un numero dispari. La tappezzeria nasce come risultato di figure di interferenza ripetitive nelle direzioni orizzontale e verticale. Può capitare che un gran numero di punti consecutivi della griglia cada all'interno degli anelli; il gruppo successivo cadrà quindi all'esterno, e così via. A mano a mano che la distanza dall'origine aumenta, gli anelli diventano progressivamente più piccoli, garantendo così che i centri e i colpi mancati avvengano lungo qualsiasi linea di punti della griglia.

Un sostegno a queste affermazioni può venire dall'esperienza di Paul Braun di Simi Valley, California, che si è fatto vincere dall'impazienza. Lo schermo del suo calcolatore ci metteva così tanto a riempirsi che egli decise di prendere un campione del disegno visualizzando un pixel ogni otto nella direzione orizzontale e in quella verticale. In forma compressa, l'immagine risultante non assomigliava affatto all'originale; Braun, in effetti, aveva cambiato la dimensione della griglia, alterando in questo modo la figura d'interferenza.

Otto Smith di Port Townsend, Washington, ha scoperto la possibilità di modificare notevolmente la configurazione moiré variando semplicemente lo schema dei colori. Come altri lettori, Smith ha scelto formule diverse da quella di Connett. Certe somme o prodotti di semplici funzioni trigonometriche, per esempio, producono immagini agitate e vorticose che ricordano i disegni a inchiostro colorato usati per abbellire le rilegature interne dei libri pregiati. Smith rileva anche che figure moiré e altri effetti di interferenza si presentano con regolarità nella grafica al calcolatore. In questo contesto sono noti sotto il nome di *aliasing*: la tendenza alla comparsa di immagini non volute quando viene digitalizzata una raffigurazione contenente variazioni regolari molto fini.

Mi ha fatto piacere ricevere un «film manuale» in miniatura, che è stato realizzato da Douglas W. Raymond di Orinda, California, ed è formato da minuscole stampe consecutive di tappezzerie di Connet in cui è gradualmente aumentata la dimensione della griglia. Le stampe sono cucite insieme e lo «spettacolo» si ottiene semplicemente facendo scorrere con il pollice le pagine del libretto risultante. Raymond intitola il suo film «Small Bang with Aliases».